Rethinking Memory in AI: Taxonomy, Operations, Topics, and Future Directions

Yiming Du^{1,2}*, Wenyu Huang²*, Danna Zheng²*, Zhaowei Wang³, Sebastien Montella⁴, Mirella Lapata², Kam-Fai Wong¹, Jeff Z. Pan^{2,4}

¹The Chinese University of Hong Kong ²The University of Edinburgh ³HKUST

⁴Poisson Lab, CSI, Huawei UK R&D Ltd.

ydu@se.cuhk.edu.hk, {w.huang, dzheng}@ed.ac.uk, https://knowledge-representation.org/j.z.pan/

Abstract

Memory is a fundamental component of AI systems, underpinning large language models (LLMs) based agents. While prior surveys have focused on memory applications with LLMs, they often overlook the atomic operations that underlie memory dynamics. In this survey, we first categorize memory representations into parametric, contextual structured, and contextual unstructured and then introduce six fundamental memory operations: Consolidation, Updating, Indexing, Forgetting, Retrieval, and Compression. We systematically map these operations to the most relevant research topics across long-term, long-context, parametric modification, and multi-source memory. By reframing memory systems through the lens of atomic operations and representation types, this survey provides a structured and dynamic perspective on research, benchmark datasets, and tools related to memory in AI, clarifying the functional interplay in LLMs based agents while outlining promising directions for future research¹.

1 Introduction

Memory is a critical component of LLM-based AI systems (Wang et al., 2024j), enabling them to sustain coherent and long-term interactions (Maharana et al., 2024; Li et al., 2024a). Although recent studies have explored various mechanisms such as memory storage (Zhong et al., 2024), memory retrieval (Qian et al., 2024; Wang et al., 2025a), and memory-grounded generation (Lu et al., 2023; Yang et al., 2024; Lee et al., 2024b)— systematic perspectives that unify these components into cohesive memory architectures are still in their early stages of development (He et al., 2024c).

Recently, a number of surveys have proposed operational views of memory (Zhang et al., 2024d)

in an attempt to organize previous works systematically. However, those surveys only focus narrowly and partially on the memory problem, such as longcontext modeling (Huang et al., 2023b), long-term memory (He et al., 2024c; Jiang et al., 2024b), personalization (Liu et al., 2025a), or knowledge editing (Wang et al., 2024g), and generally lack a unified and comprehensive view of memory operations, as well as in-depth discussions of technical pathways. For example, Zhang et al. (2024d) cover only high-level operations such as writing, management, and reading and miss some operations like indexing. In addition, most existing surveys do not clarify the overall scope of memory research in AI and overlook practical foundations for further work—such as structured benchmark categorization and coverage of relevant tools.

To address these gaps, our survey divides memory into three types: parametric memory (Wang et al., 2024c), contextual-structured memory (Rasmussen et al., 2025), and contextual-unstructured memory (Zhong et al., 2024). Drawing inspiration from cognitive psychology, we further classify these memory types by their temporal span: long-term memory (i.e., persistent memory such as multi-turn dialogue history (Zhong et al., 2024)), external environment observations (Li et al., 2024a), or internal parameters (Wang et al., 2024c), and short-term memory corresponds to transient, session-level contexts (Packer et al., 2023).

Based on the memory types, we classify memory operations into six fundamental types, spanning both memory management and utilization. Specifically, memory management comprises four operations: **consolidation** (integrating new knowledge into persistent memories (Feng et al., 2024)), **indexing** (efficiently organizing memory for retrieval (Wu et al., 2024a)), **updating** (modifying memory in response to new data (Chen et al., 2024b)) and **forgetting** (strategically removing outdated or less relevant memories (Tian et al., 2024)). Mem-

^{*}These authors contributed equally.

¹The paper list, datasets, methods and tools are available at https://github.com/Elvin-Yiming-Du/Survey_Memory_in_AI.

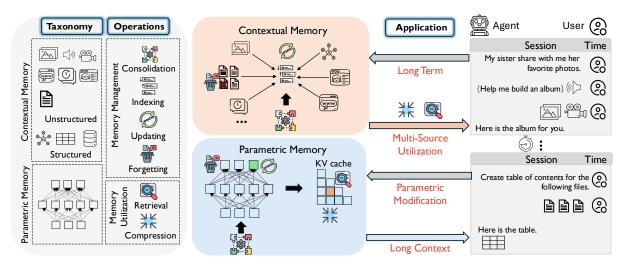


Figure 1: A unified framework of memory Taxonomy, Operations, and Applications in AI systems.

ory utilization involves two operations: **retrieval** (accessing relevant memory content when needed (Gutiérrez et al., 2024)) and **compression** (reducing memory size while preserving essential information for efficient storage and reasoning (Chen et al., 2024b)).

To better ground this taxonomy and map the research landscape, we conduct a comprehensive literature analysis. Prior surveys (Zhang et al., 2024d; He et al., 2024c; Jiang et al., 2024b) lack a clearly defined scope of the collected literature, making it difficult to assess impactful memory-related research or to chart future directions for memory-centric LLM-based agents. In response, we collect and review more than 30K top-tier conference papers² published between 2022-2025. To identify influential memory-focused studies, we apply the Relative Citation Index (RCI), which ranks papers by normalized citation impact over time (see Appendix A). This analysis reveals four primary research topics critical to memory in AI systems:

- Long-Term Memory, focusing on memory management, inference, and personalization in multi-session dialogue systems (Xu et al., 2021; Maharana et al., 2024), retrieval-augmented generation (RAG), personalized agents (Li et al., 2024a), and question answering tasks (Wu et al., 2024a; Zhong et al., 2024).
- Long-Context Memory, addressing both parametric efficiency (e.g. "KV cache dropping" (Zhang et al., 2023d)) and context uti-

lization effectiveness (e.g., long-context compression (Cheng et al., 2024; Jiang et al., 2024a)) in handling extended sequences.

- Parametric Memory Modification, focused on model editing (Fang et al., 2024; Meng et al., 2022b; Wang et al., 2024c), unlearning (Maini et al., 2024), and continual learning (Wang et al., 2024j) for adapting internal knowledge representations.
- Multi-Source Memory, emphasizing integration across heterogeneous textual sources (Li et al., 2023; Hu et al., 2023) but also multimodal inputs (Wang et al., 2025a) to further support robust and scene-awareness reasoning.

The remainder of the paper is organized as follows. Section 2 introduces the memory taxonomy, including a temporal perspective distinguishing short-term and long-term memory. Section 3 presents key memory operations, clarifying how memory is managed and utilized across different research contexts. Section 4 maps representative research topics to memory taxonomies and operations, highlights emerging trends based on high-RCI papers (RCI > 1), and summarizes key methods and datasets (see Appendix 6). We also include some topic recommendations at the end of each sub-section. In Section 5, we introduce a suite of practical tools, including components, frameworks, application layers, and real-world products supporting memory integration into modern AI applications. In Section 6, we conclude by outlining long-term visions for the future of memory in AI.

²Across NeurIPS, ICLR, ICML, ACL, EMNLP, and NAACL conferences.

2 Memory Taxonomy

From the perspective of memory representation, we outline three representative types of memory: Parametric Memory, Contextual Unstructured Memory, and Contextual Structured Memory.

Parametric Memory refers to the knowledge implicitly stored within a model's internal parameters (Berges et al., 2024; Wang et al., 2024c; Prashanth et al., 2024). Acquired during pretraining or post-training, this memory is embedded in the model's weights and accessed through feedforward computation at inference time. It serves as a form of instant, long-term, and persistent memory enabling fast, context-free retrieval of factual and commonsense knowledge. However, it lacks transparency and is difficult to update selectively in response to new experiences or task-specific contexts.

Contextual Unstructured Memory refers to an explicit, modality-general memory system which stores and retrieves information across heterogeneous inputs such as text (Zhong et al., 2024), images (Wang et al., 2025a), audio, and video (Wang et al., 2023b). It enables agents to ground reasoning in perceptual signals and integrate multi-modal context (Li et al., 2024a). Depending on its temporal scope, it is further divided into short-term and long-term. Short-term memory refers to the recent observations like the current dialogue session context, while long-term memory refers to the persistent records of cross-session conversation dialogues and personal persistent knowledge.

Contextual Structured Memory denotes an explicit memory organized into predefined, interpretable formats or schemata such as knowledge graphs (Oguz et al., 2022), relational tables (Lu et al., 2023), or ontologies (Qiang et al., 2023) while remaining easily queryable upon requests. These structures support symbolic reasoning and precise querying, often complementing the associative capabilities of pretrained language models (PLMs). Structured memory can be short-term, constructed at inference for local reasoning, or long-term, storing curated knowledge across sessions.

3 Memory Operations

To enable dynamic memory beyond static storage, AI systems require operations that govern the lifecycle of information and support its effective use during interaction with the external environment. These operations can be grouped into two functional categories: Memory Management and Memory Utilization.

3.1 Memory Management

Memory management governs how memory is stored, maintained, and pruned over time. It includes four core operations: Consolidation, Indexing, Updating, and Forgetting. These operations naturally incorporate the temporal nature of memory, where information evolves over time.

Consolidation (Squire et al., 2015) refers to transforming short-term experiences into persistent memory. This involves encoding interaction histories (i.e. dialogs, trajectories, etc.) into durable forms such as model parameters (Wang et al., 2024j), graphs (Zhao et al., 2025), or knowledge bases (Lu et al., 2023). It is essential for continual learning (Feng et al., 2024), personalization (Zhang et al., 2024a), external MemoryBank construction (Zhong et al., 2024), and knowledge graph construction (Xu et al., 2024c).

Indexing (Maekawa et al., 2023) refers to the construction of auxiliary codes—such as entities, attributes, or content-based representations (Wu et al., 2024a)—that serve as access points to stored memory. Beyond simple access, indexing also enables the encoding of temporal (Maharana et al., 2024) and relational structures (Mehta et al., 2022) across memories, allowing for more efficient and semantically coherent retrieval through traversable index paths. It supports scalable retrieval across symbolic, neural, and hybrid memory systems.

Updating (Kiley and Parks, 2022) reactivates existing memory representations and temporarily modify them. Updating parametric memory typically involves a locate-and-edit mechanism (Fang et al., 2024) that targets specific model components. Meanwhile, contextual memory updating involves summarization (Zhong et al., 2024), pruning, or refinement (Bae et al., 2022) to reorganize or replace outdated content. Those updating operations support continual adaptation while maintaining memory consistency.

Forgetting (Davis and Zhong, 2017) is the ability to selectively suppress memory content that may be outdated, irrelevant or harmful. In parametric memory, it is commonly implemented through unlearning techniques (Jia et al., 2024a;

Li et al., 2025) that modify model parameters to erase specific knowledge. In contextual memory, forgetting involves time-based deletion (Zhong et al., 2024) or semantic filtering (Wang et al., 2024e) to discard content that is no longer relevant. These operations help maintain memory efficiency and reduce interference.

However, these operations introduce inherent risks and limitations. Attackers can exploit vulnerabilities to alter or poison memory contents. Once corrupted, memory fragments may persist undetected and later trigger malicious actions. As discussed in Section 6, such threats call for robust approaches that address not only the memory operations but also the entire memory lifecycle.

3.2 Memory Utilization

The way in which stored memory is retrieved and used during inference is referred as memory utilization. It includes two operations: retrieval and compression.

Retrieval is the process of identifying and accessing relevant information from memory in response to inputs, aiming to support downstream tasks such as response generation, visual grounding, or intent prediction. Inputs can range from a simple query (Du et al., 2024) to a complex multi-turn dialogue context (Wang et al., 2025a), and from purely textual inputs to visual content (Zhou et al., 2024) or even more modalities. Retrieval targets include memory from multiple sources (Tan et al., 2024a), modalities (Wang et al., 2025a), or even parametric representations (Luo et al., 2024) within models.

Compression enables efficient context usage under limited context window by retaining salient information and discarding redundancies before feeding it into models. It can be broadly divided into preinput compression and post-retrieval compression. Pre-input compression applies in long-context models without retrieval, where full-context inputs are scored, filtered, or summarized to fit within context constraints (Yu et al., 2023; Chung et al., 2024). Post-retrieval compression operates after memory access, reducing retrieved content either through contextual compression before model inference (Xu et al., 2024a) or through parametric compression by integrating retrieved knowledge into model parameters (Safaya and Yuret, 2024). Unlike memory consolidation, which summarizes information during memory construction (Zhong et al., 2024), compression focuses on reducing memory at inference (Lee et al., 2024b).

4 From Operations to System-Level Topics

Building on the core operations introduced above, this section examines how real-world systems coordinate these operations to support complex memory usage patterns. For instance, many cross-session dialogue systems (Maharana et al., 2024; Wu et al., 2024a) adopt RAG frameworks where updating, indexing, retrieval, and compression work in concert to generate robust responses. Meanwhile, many approaches (Packer et al., 2023) treat long-term memory as long-context inputs (Packer et al., 2023; Xiao et al., 2024), where retrieval and compression are essential.

We categorize combined memory usages into four key topics derived from memory types (Figure 2): long-term (all types), long-context (contextual-unstructured), parametric modification (parametric), and multi-source memory (contextual). The mapping between different memory operations and types under each topic is summarized in Table 1. Each section concludes with topic-specific insights on current limitations and future directions. The Appendix complements the survey with representative methods, benchmarks, and curated datasets—annotated with key technical details, memory operations, evaluation protocols, and usage scenarios.

4.1 Long-term Memory

Long-term memory refers to the persistent storage of information acquired through interactions with the environment, such as multi-turn dialogues, browsing patterns, and agent decision paths. It supports capabilities such as memory management, utilization, and personalization over extended interactions, enabling agents to perform complex tasks over time. We review representative datasets addressing long-term memory processing and personalization (see Table 2). This section focuses on contextual long-term memory-structured or unstructured—which differs from parametric memory stored in model weights via continual learning and memory editing. Expanded summaries of datasets and methods are provided in Appendix Tables 2 and 6.

Operations	Parametric	Contextual-Structured	Contextual-Unstructured
Consolidation	Continual Learning, Personalization	Management, Personalization	Management, Personalization
Indexing	Utilization	Utilization, Management, Multi-modal Coordination	Utilization, Management, Personalization
Updating	Knowledge Editing	Cross-Textual Integration, Personalization, Management	Cross-Textual Integration, Personalization, Management
Forgetting	Knowledge Unlearning, Personalization	Management	Management
Retrieval	Utilization, Parametric Efficiency	Utilization, Personalization, Contextual Utilization, Multi-modal Coordination	Utilization, Personalization, Contextual Utilization
Compression	Parametric Efficiency	Contextual Utilization	Contextual Utilization

Table 1: Alignment of sub-topics with memory types and memory operations. Sub-topics are highlighted with colors with respect to the topics: Long-term, Long-context, Parametric, Multi-source.

4.1.1 Management

Management in long-term memory involves operations such as consolidation, indexing, updating, and forgetting of acquired experiences. Here, memory is instantiated in two forms: (1) accumulated dialogue histories from multi-turn conversations, and (2) long-term observations and decisions made by autonomous agents. These are often encoded by LLMs and stored in external memory repositories for future access and reuse. The memory in those tasks is routinely updated with new information and pruned to remove outdated or irrelevant content.

Memory Consolidation refers to the process of transforming short-term memory into long-term memory. This often involves saving dialogue history into persistent memory. Existing approaches commonly adopt summarization techniques to generate unstructured memory representations, as seen in systems like MemoryBank (Zhong et al., 2024) or ChatGPT-RSum (Wang et al., 2025c). To facilitate the extraction of key topics and salient memory elements, Lu et al. (2023) utilize LLM prompting to identify and structure relevant information. Different from summarization, MyAgent (Hou et al., 2024) emphasizes context-aware memory strengthening by modeling temporal relevance. Beyond dialogue agent task-based systems (Park et al., 2025) incorporate episodic what-where-when memories to hierarchically organize long-term knowledge for action planning. Together, these works illustrate a growing effort to integrate human-like memory consolidation processes into LLM-based agents.

Memory Indexing refers to the process of structuring memory representations to support efficient and accurate retrieval since standing as a foundational component of memory usage. Recent work categorizes memory indexing into three paradigms: graph-based, signal-enhanced, and timeline-based approaches. HippoRAG (Gutiérrez et al., 2024) models memory indexing after hippocampal theory by constructing lightweight knowledge graphs to explicitly reveal the connection between different knowledge fragments. LongMemEval (Wu et al., 2024a) enhances memory keys with timestamps, factual content, and summaries. Theanine (iunn Ong et al., 2025) organizes memories along evolving temporal and causal links, enabling dialogue agents to retrieve information segments based on both relevance and timeline context, supporting lifelong and dynamic personalization. These strategies highlight the need to integrate structure, retrieval signals, and temporal dynamics for effective long-term memory management.

Memory Updating typically denotes the process by which external memory either creates new entries for unseen information (Chen et al., 2024b), or reorganizes and integrates content with existing memory representations (Bae et al., 2022). Recent research categorizes memory updating into two overarching paradigms: intrinsic updating and extrinsic updating. Intrinsic Updating operates through internal mechanisms without explicit external feedback. Techniques such as selective editing (Bae et al., 2022) manage memory by selectively deleting outdated information, while recursive summarization (Wang et al., 2025b) compresses dialogue histories through iterative summar

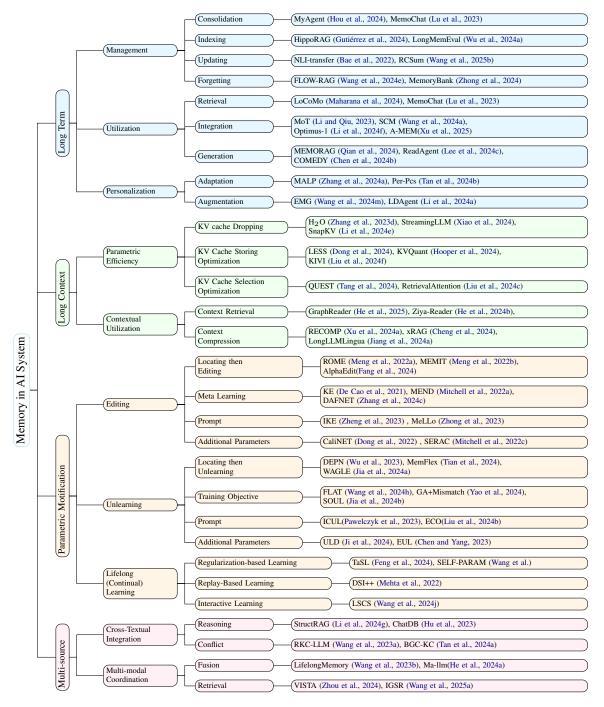


Figure 2: Operation-driven system-level topics in AI systems.

rization. Memory blending and refinement (Kim et al., 2024b) further evolve memory by merging past and present representations, and self-reflective memory evolution (Sun et al., 2024) updates memory based on evidence retrieval and verification, enhancing factual consistency over time. **Extrinsic Updating** relies on external signals, particularly user feedback. For instance, dynamic feedback incorporation (Dalvi Mishra et al., 2022) stores user corrections into memory, enabling continual system improvement without requiring retrain-

ing. These approaches emphasize balancing selforganized memory updates and user-driven adaptations for scalable long-term memory.

Memory Forgetting involves the removal of previously consolidated long-term memory representations. Forgetting can occur naturally over time, for example, following the Ebbinghaus forgetting curve (Zhong et al., 2024), where memory traces decay gradually. In contrast, active forgetting strategies (Chen et al., 2024b; Mitchell et al.,

2022c) have been developed to intentionally remove specific information from memory systems. This is particularly important when long-term memory stores sensitive or potentially harmful content. Therefore, enabling systems to intentionally remove specific content for reasons such as privacy, safety, or compliance has become a major focus (Liu et al., 2024e; Eldan and Russinovich, 2024; Ji et al., 2024; Li et al., 2025; Liu et al., 2025b).

4.1.2 Utilization

Utilization refers to the process of generating responses conditioned on current inputs and relevant memory content, typically involving memory routing, integration, and reading.

Memory Retrieval focuses on the selection of the most relevant memory entries based on a given query. To systematize recent progress, retrieval methods can be broadly categorized into three paradigms: (1) query-centered retrieval, which focuses on improving query formulation and adaptation, such as forward-looking query rewriting in FLARE (Jiang et al., 2023b) and iterative refinement in IterCQR (Jang et al., 2024); (2) memorycentered retrieval, which enhances the organization and ranking of memory candidates, including better indexing strategies (Wu et al., 2024a) and reranking methods (Du et al., 2024); and (3) event-centered retrieval, which retrieves memories based on temporal and causal structures, as explored in LoCoMo (Maharana et al., 2024), CC (Jang et al., 2023) and MSC (Xu et al., 2021). Other techniques, such as multi-hop graph traversal (Gutiérrez et al., 2024) and memory graph evolution (Qian et al., 2024), further enrich the retrieval process. These approaches highlight the importance of adaptive retrieval for effective long-term memory access, although reasoning over evolving memory sequences remains an open challenge.

Memory Integration refers to the process of selectively combining retrieved memory with the model context to enable coherent reasoning or decision-making during inference. Integration may span multiple memory sources (e.g., long-term dialogue histories, external knowledge bases) and modalities (e.g., text, images, or videos), enabling richer and contextually grounded generation. Recent efforts on memory integration can be broadly categorized into two strategies. Static contextual integration approaches, such as EWE (Chen et al., 2024a) and Optimus-1 (Li et al., 2024f), focus

on retrieving and combining static memory entries at inference time to enrich context and improve reasoning consistency. In contrast, **dynamic memory evolution** approaches, exemplified by A-MEM (Hou et al., 2024), Synapse (Zheng et al., 2024), R2I (Samsami et al., 2024) and SCM (Wang et al., 2024a), emphasize enabling memory to grow, adapt, and restructure over the course of interactions, either through dynamic linking or controlled memory updates. While static integration enhances immediate contextual grounding, dynamic evolution is crucial for building more adaptive, lifelong learning agents.

Memory Grounded Generation refers to utilizing retrieved memory content that has been integrated to guide the generation of responses. Existing methods can be broadly categorized into three types based on how memory influences generation. First, Self-Reflective Reasoning methods, such as MoT (Li and Qiu, 2023) and StructRAG (Li et al., 2024g), retrieve self-generated or structured memory traces to guide intermediate reasoning steps, enhancing multi-hop inference during decoding. Second, Feedback-Guided Correction approaches, including the method of MemoRAG (Qian et al., 2024) and Repair (Tandon et al., 2021), leverage feedback memories or memory-informed clues to constrain generation, preventing repeated errors and improving output robustness. Third, **Contextually-Aligned Long-Term Generation** techniques, exemplified by COMEDY (Chen et al., 2024b), MemoChat (Lu et al., 2023), and ReadAgent (Lee et al., 2024c), integrate compressed or extracted memory summaries into the generation process to maintain coherence over long dialogues or extended documents. These methods collectively enhance generation quality, consistency, and reasoning depth, though challenges like noise in memory and reliability of retrieved memories remain to be addressed.

4.1.3 Personalization

Personalization is a key but challenging aspect of long-term memory, constrained by data sparsity, privacy, and changing user preferences. Current methods can be broadly categorized into two lines: model-level adaptation and memory-level augmentation.

Model-Level Adaptation encodes user preferences into model parameters via fine-tuning or lightweight updates. Some methods embed user

traits in latent space—e.g., CLV (Tang et al., 2023a) uses contrastive learning to cluster persona descriptions for guiding generation. Others adopt parameter-efficient strategies: RECAP (Liu et al., 2023b) injects retrieved user histories via a prefix encoder, while Per-Pes (Tan et al., 2024b) assembles modular adapters to reflect user behaviors. In specialized domains, MaLP (Zhang et al., 2023c) introduces a dual-process memory for modeling short- and long-term personalization in medical dialogues. These methods show how lightweight adaptation can personalize models without compromising efficiency or generalizability.

Memory-Level Augmentation personalizes LLMs by retrieving user-specific information from external memory at inference time. Based on the memory format, existing methods can be categorized into structured, unstructured, and hybrid approaches. Structured memories, such as user profiles or knowledge graphs, are used in LaMP (Salemi et al., 2023) to construct personalized prompts and in PerKGQA (Dutt et al., 2022) for question answering over individualized subgraphs. Unstructured memories, including dialogue histories and narrative personas, are retrieved in LAPDOG (Huang et al., 2024) to enrich sparse profiles and aligned with input contexts via dual learning in Fu et al. (2022). Hybrid methods like SiliconFriend (Zhong et al., 2024) and LD-Agent (Li et al., 2024a) maintain persistent memory across sessions. While these approaches demonstrate scalability, they often treat long-term memory as a passive buffer, leaving its potential for proactive planning and decision-making underexplored.

Long-term assistants mainly extract or summarize memory, but rarely model the evolution of personalized memory with changing user interests and external knowledge.

Time-indexed memory improves retrieval, but its use for long-term reasoning remains limited.

While personalization leverages longterm memory, its reuse and integration into memory-guided planning remain underexplored.

4.2 Long-context

Managing vast quantities of multi-sourced external memory in conversational search presents significant challenges in long-context language understanding. While advancements in model design and long-context training have enabled LLMs to process millions of input tokens, effectively managing memory within such extensive contexts remains a complex issue. These challenges can be broadly categorized into two main aspects: 1) Parametric Efficiency, which focuses on optimizing the KV cache (parametric memory) to enable efficient long context decoding and Contextual Utilization optimizes the utilization of LLMs to manage various external memory (contextual memory). In this section, we systematically review efforts made in handling these challenges. A detailed overview of relevant datasets are discussed in Table 3, while an expand summary of highlighted works are discussed in Table 8 and Table 9.

4.2.1 Parametric Efficiency

To manage extensive amounts of multi-sourced external memory, Large Language Models (LLMs) must be optimized to efficiently process lengthy contexts. In this section, we discuss approaches for efficiently processing long-context from memory perspective, which focuses on Key-Value (KV) cache optimization. KV cache aims to minimize unnecessary key-value computations by storing past key-value pairs as external parametric memory. However, as context length increases, the memory requirement for storing these memory grows quadratically, making it infeasible for handling extremely long contexts.

KV Cache Dropping aims to reduce cache size by eliminating unnecessary KV cache. Static dropping approaches select unnecessary cache with fixed pattern. For instance, StreamingLLM (Xiao et al., 2024) and LM-Infinite (Han et al., 2024) use an Λ -shaped sparse pattern. In contrast, dynamic dropping approaches are more flexible, which decide the KV cache to be eliminated with respect to the query (e.g., H₂O (Zhang et al., 2023d), Fast-Gen (Ge et al., 2024), Keyformer (Adnan et al., 2024), Radar (Hao et al., 2025), NACL (Chen et al., 2024c)) or the model behavior (attention weight) during inference (e.g., SnapKV (Li et al., 2024e), HeadKV (Fu et al., 2025), Scissorhands (Liu et al., 2023d)). KV cache dropping methods prevent the quadratic grow of memory cost for storing KV

cache, but also introduce the risk of potential information loss when discarding KV cache.

KV Cache Storing Optimization considers the potential information loss when removing less important elements, and focus on how to preserve the entire KV cache at a smaller footprint. For instance, LESS (Dong et al., 2024) compress less important cache entries into low-rank representations, while FlexGen (Sheng et al., 2023), Atom (Zhao et al., 2024c), KVQuant (Hooper et al., 2024) and KIVI (Liu et al., 2024f) dynamically quantize KV cache to reduce memory allocation. These approaches provide less performance drop compared with KV cache dropping methods but remain limited due to the quadratic nature of the growing memory. Future works should continue focusing on the tradeoff between less memory cost and less performance drop.

KV Cache Selection refers to selectively loading required KV cache to speed up the inference, which focus on memory retrieval upon KV cache. QUEST (Tang et al., 2024) and TokenSelect (Wu et al., 2025) adapt query-aware KV cache selection to retrieve critical KV cache for accelerate inference. Similarly, RetrievalAttention (Liu et al., 2024c) adopts Approximate Nearest Neighbor (ANN) to search critical KV cache. These methods offer greater flexibility as they avoid evicting the KV cache and have the potential to integrate with storage optimization techniques (e.g., Tang et al. (2024) shows QUEST is compatible with Atom (Zhao et al., 2024c)).

4.2.2 Contextual Utilization

Apart from optimizing language models to obtain long-context abilities, optimizing memory utilization raises another important challenge. Despite claims that context length can extend to millions of tokens, long-context LLMs have been found to miss crucial information in the middle of the context during tasks such as question answering and key-value retrieval (Liu et al., 2024d; Ravaut et al., 2024). This "lost in the middle" issue is especially critical when managing vast amounts of external memory, as essential information may be located at various positions within the long context. In addition, though higher recall can be obtained with larger retrieval set, the hard-negative information will mislead LLMs and harm the generation quality (Jin et al., 2025). Effective contextual utilization become a key challenge in addressing these limitations, encompassing context retrieval and context compression across memory operations.

Context Retrieval aims to enhance LLM's ability in identifying and locating key information from the contextual memory. Graph-based approaches like CGSN (Nie et al., 2022) and GraphReader (Li et al., 2024c)) decompose documents into graph structure for effective context selection. Tokenlevel context selection approaches (e.g., TRAMS (Yu et al., 2023), Selection-p (Chung et al., 2024)) assign scores to individual tokens, pruning and selecting those deemed most important. In contrast, methods such as NBCE (Su et al., 2024), FragRel (Yue et al., 2024), and Sparse RAG (Zhu et al., 2025) perform context selection at the fragment level, choosing the relevant context fragments based on their importance to the specific task. Furthermore, training-based approaches like Ziya-Reader (He et al., 2024b) train LLMs with specialized data to help improve their context selection ability. Other methods like Neurocache (Safaya and Yuret, 2024) and AWESOME (Cao and Wang, 2024) preserve an external vector memory cache to effectively store and retrieve first encode external memory into vector space, and this external vector memory can be effectively updated or retrieved to enable long-term memory utilization. Together with these methods, LLMs are allowed to better identify key information in the context via memory retrieval.

Context Compression utilizes memory compression operation to optimize contextual memory utilization, which generally involves two major approaches: soft prompt compression and hard prompt compression (Li et al., 2024h). Soft prompt compression focuses on compressing chunks of input tokens into the continuous vectors to reduce the input sequence length. For instance, xRAG (Cheng et al., 2024), which use a off-the-shelf sentence encoder plus a trained projector to convert context document into a document embedding. While hard prompt compression directly compress long input chunks into shorter natural language chunks. Lee et al. (2024c) utilize LLM to shorten context page into natural language memory gist. Similarly, Xu et al. (2024a) introduce RECOMP composed with two compressor to summarize retrieved documents and reduce context length. Meanwhile, Jiang et al. (2023a) and Jiang et al. (2024a) introduce iterative token-level prompt compression to compress

prompt to support long memory utilization. With both soft prompt and hard prompt, LLMs are allowed to more effectively utilize the context via memory compression.

Balancing the trade-off between reduced memory usage and minimized performance degradation in KV cache optimization represents an exciting area for future research.

Contextual utilization with complex environment (e.g., multi-source memory) is a pivotal research direction for advancing the development of intelligent agents.

4.3 Parametric Memory Modification

Modifying parametric memory, which refers to knowledge encoded within the parameters of LLMs, is crucial for dynamically adapting stored memory. Methods for parametric memory modification can be broadly categorized into three types: (1) **Editing** refers to the localized modification of model parameters without requiring full model retraining; (2) Unlearning, which selectively removes unwanted or sensitive information; and (3) Continual Learning, which incrementally incorporates new knowledge while mitigating catastrophic forgetting. This section systematically reviews recent research in these categories, with detailed analyses and comparisons presented in subsequent subsections. A comprehensive overview of relevant datasets is presented in Table 4, and extended summaries of key methods are provided in Tables 10, Table 11 and Table 12.

4.3.1 Editing

Parametric memory editing updates specific knowledge stored in the parametric memory without full retraining. The predominant approach is locatingthen-editing method (Meng et al., 2022a, 2023; Mela et al., 2024; Fang et al., 2025), which uses attribution or tracing to find where facts are stored, then modifies the identified memory directly. Another major method is meta-learning (De Cao et al., 2021; Mitchell et al., 2022b; Zhang et al., 2024c), where an editor network learns to predict targeted weight changes for quick and robust corrections. Other strategies avoid editing the original parameters. Prompt-based methods (Zheng et al., 2023; Zhong et al., 2023) use crafted prompts like ICL to steer outputs indirectly. Additional-parameter methods (Wang et al., 2024c; Dong et al., 2022; Mitchell et al., 2022c; Wang et al., 2024i; Das et al., 2024) add external parametric memory modules to adjust behavior without touching model weights. These approaches vary in efficiency and scalability, though most focus on entity-level edits.

4.3.2 Unlearning

Parametric memory unlearning enables selective forgetting by removing specific memory while retaining unrelated memory. Recent work explores several strategies. Additional-parameter methods add components such as logit difference modules (Ji et al., 2024) or unlearning layers (Chen and Yang, 2023) to adjust memory without retraining the whole model. Prompt-based methods manipulate inputs (Liu et al., 2024b) or use ICL (Pawelczyk et al., 2024) to externally trigger forgetting. Locating-then-unlearning methods (Jia et al., 2024a; Tian et al., 2024; Wu et al., 2023) first identify responsible parametric memory, then apply targeted updates or deactivations. Training objective-based methods (Wang et al., 2025d; Jia et al., 2024b; Yao et al., 2024) modify the training loss functions or optimization strategies explicitly to encourage memory forgetting. These approaches aim to erase memory when given explicit forgetting targets, while preserving non-targeted knowledge and balancing efficiency and precision.

4.3.3 Continual Learning

Continual learning (Wang et al., 2024b) enables long-term memory persistence by mitigating catastrophic forgetting in model parameters. Two main approaches are regularization-based and replaybased methods. Regularization constrains updates to important weights, preserving vital parametric memory; methods like TaSL (Feng et al., 2024), SELF-PARAM (Wang et al.), EWC (Kirkpatrick et al., 2017), and POCL (Wu et al., 2024b) apply such constraints to embed knowledge without replay. In contrast, replay-based methods reinforce memory by reintroducing past samples, and are particularly suited to incorporating retrieved external knowledge or historical experiences during training. For example, DSI++ (Mehta et al., 2022) leverages generative memory to supplement learning with pseudo queries, maintaining retrieval performance without full retraining. Beyond these paradigms, agent-based work such as LifeSpan Cognitive System (LSCS) (Wang et al., 2024j) extends continual learning into an interactive setting, enabling agents to incrementally acquire and consolidate memory

through real-time experience. LSCS provides valuable insights into how external memory can be encoded into model parameters in a continual manner.

Current editing methods mainly focus on entity replacement but lack mechanisms to model broader memory evolution following diverse knowledge updates.

Current unlearning methods typically target specific sequences, but future challenges involve erasing all parametric memories linked to given keywords without needing explicit content specification.

Current agents accumulate memory through interaction, but future continual learning should avoid overwriting persistent memory in model parameters.

4.4 Multi-source Memory

Multi-source memory is essential for real-world AI deployment, where systems must reason over internal parameters and external knowledge bases, spanning structured data (e.g., knowledge graphs, tables) and unstructured multi-modal content (e.g., text, audio, images, videos). This section examines key challenges across two dimensions: cross-textual integration and multi-modal coordination. A detailed overview of the datasets and an expanded summary of the methods discussed are provided in Appendix Table 5, Table 13 and Table 14, respectively.

4.4.1 Cross-textual Integration

Cross-textual integration enables AI systems to perform deeper reasoning and resolve conflicts from multiple textual sources to support more contextually grounded responses.

Reasoning focuses on integrating multi-format memory to generate factually and semantically consistent responses. One line of research investigates reasoning over memories from different domains, particularly through the precise manipulation of structured symbolic memories, as demonstrated by ChatDB (Hu et al., 2023) and Neurosymbolic (Wang et al., 2024f). Other works (Nogueira dos Santos et al., 2024; Wu et al., 2022) explore the dynamic integration of domain-specific parameterized memories to enable more flexible reasoning. Multi-source reasoning across diverse doc-

ument sources has also been studied, as seen in DelTA (Wang et al., 2025e) and dynamic-MT (Du et al., 2022). Additionally, several studies (Li et al., 2024g; Lee et al., 2024a; Zhao et al., 2024b; Xu et al., 2024c) have investigated heterogeneous knowledge integration by retrieving information from both structured and unstructured sources. Despite progress in combining parameterized and external memories, unified reasoning over heterogeneous, multi-source knowledge remains a major challenge, particularly in integrating parameterized memory with both structured and unstructured sources.

Conflict in multi-source memory refers to factual or semantic inconsistencies that arise during the retrieval and reasoning over heterogeneous memory representations by AI systems. In memoryaugmented architectures, conflicts typically emerge during the integration of memory representations originating from heterogeneous sources, including parametric and contextual memories as well as structured and unstructured knowledge such as triples, tables, and free text (Xu et al., 2024b). Existing work (Wang et al., 2023a; Tan et al., 2024a) has primarily focused on the identification and localization of conflicts. For example, RKC-LLM (Wang et al., 2023a) proposes an evaluation framework for assessing models' ability to detect and localize contextual contradictions, while BGC-KC (Tan et al., 2024a) reveals a systematic bias toward internal knowledge over external sources, underscoring the need for trust calibration and source attribution. However, resolving memory conflicts remains an open challenge, requiring not only factual verification but also the semantic alignment of memories across structurally and temporally heterogeneous sources.

4.4.2 Multi-Modal Coordination.

As memory-augmented systems evolve toward multi-modal settings, a key challenge lies in fusion and retrieval over heterogeneous modalities such as text, image, audio and video.

Fusion refers to aligning the retrieved information across diverse modalities. From a memory perspective, fusion serves as a key mechanism for integrating cross-modal information over time. Existing approaches can be broadly into two lines. The first focuses on **unified semantic projection**, where models such as UniTransSeR (Ma et al., 2022), MultiInstruct (Xu et al., 2023), PaLM-E

(Driess et al., 2023), and NExT-Chat (Zhang et al., 2023a) embed heterogeneous inputs into a shared representation space for reuse and query. The second line emphasizes long-term cross-modal memory integration. For example, LifelongMemory (Wang et al., 2023b) introduces a transformer with persistent memory to accumulate visual-textual knowledge across patient records. Similarly, MA-LMM (He et al., 2024a) maintains a multimodal memory bank to extend temporal understanding in long videos. While effective at aligning modalities, current fusion methods often fall short in long-term multi-modal memory management. Key challenges include dynamic updates and maintaining consistency across heterogeneous sources.

Retrieval in multi-modal systems enables access to stored knowledge across modalities such as text, image, and video. Most existing methods rely on embedding-based similarity computation, grounded in vision-language models like QwenVL (Bai et al., 2023), CLIP (Radford et al., 2021) or other multi-modal models (Li et al., 2024d). These models project heterogeneous inputs into a shared semantic space, allowing for cross-modal retrieval. For instance, VISTA (Zhou et al., 2024) enhances retrieval via visual token representations, while UniVL-DR (Liu et al., 2023c) integrates video and language through a unified dual encoder. More recently, IGSR (Wang et al., 2025a) extends retrieval to multi-session conversations by introducing intent-aware sticker retrieval, though it remains anchored in similarity-based retrieval. However, these methods are limited to shallow embedding similarity and lack memory-based, reasoningaware retrieval. Modalities like audio and sensorimotor signals remain underexplored, despite their role in grounding and long-term interaction in embodied, multi-turn settings.

Joint reasoning over short-term, structured, and unstructured long-term memory remains an open and relatively underexplored challenge.

Premporal consistency conflicts in memory remain understudied, despite the intrinsic connection between memory and time.

Multimodal memory remains relatively underexplored, particularly in retrieval alignment and cross-modal representation.

5 Tools

A layered ecosystem of memory-centric AI systems has emerged to support long-term context management, user modeling, knowledge retention, and adaptive behavior. This ecosystem spans four tiers: foundational **components** (e.g., vector stores, LLMs, retrievers), modular **frameworks** for memory operations, **memory layer** systems for orchestration and persistence, and end-user-facing **products**.

Components. Foundational components provide the infrastructure upon which memory-centric systems are built. These include vector databases such as FAISS (Douze et al., 2024), graph databases like Neo4j (Neo4j, 2012), and large language models (LLMs) such as Llama (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), and DeepSeek (Liu et al., 2024a). Retrieval mechanisms—including BM25 (Robertson et al., 1995), Contriever (Izacard et al., 2021), and OpenAI embeddings (OpenAI, 2025)—enable semantic access to external memory. These components serve as the computational substrate for building memory capabilities such as grounding, similarity search, and long-context understanding.

Frameworks. On top of core infrastructure, frameworks offer modular interface for memory-related operations. Examples include Graphiti (He et al., 2025), LlamaIndex (Liu, 2022), LangChain (Chase, 2022), LangGraph (Inc., 2025), EasyEdit (Wang et al., 2024d), CrewAI (Duan and Wang, 2024), and Letta (Packer et al., 2023). These frameworks abstract complex memory processes into configurable pipelines, enabling developers to construct multi-modal, persistent, and updatable memory modules that interact with LLM agents.

Memory Layer Systems. These systems operationalize memory as a service layer, providing orchestration, persistence, and lifecycle management. Tools like Mem0 (Taranjeet Singh, 2024), Zep (Rasmussen et al., 2025), Memary (kingjulio8238, 2025), and Memobase (kingjulio8238, 2025) focus on maintaining temporal consistency, indexing memory by session or topic, and ensuring efficient recall. These platforms often combine symbolic and sub-symbolic memory representations and provide internal APIs for memory access and manipulation over time.

Products. At the application layer, memoryenabled AI is being deployed in user-facing systems that emphasize personalization, user state retention, and lifelong learning. Examples include Me.bot³, Tencent ima.copilot⁴, Coze (Coze, 2024), Grok (xAI, 2023), and ChatGPT (OpenAI, 2022). These products demonstrate how structured memory pipelines can enhance user engagement, enable long-term dialog continuity, and support user-specific reasoning across interactions.

The more details are shown in tables: Table 15 (Components), Table 16 (Frameworks), Table 17 (Memory Layer Systems), and Table 18 (Products). Each table describes the tool's applicable memory type, supported operations, input/output formats, core functionality, usage scenarios, and source type.

6 Open Challenges and Future Directions

Spatio-temporal Memory captures not only the structural relationships among information but also their temporal evolution, enabling agents (Lei et al., 2025) to adaptively update knowledge while preserving historical context (Zhao et al., 2025). For example, an AI system may record that a user once disliked broccoli but later adjust its memory based on recent purchase patterns. By maintaining access to both historical and current states, spatio-temporal memory supports temporally informed reasoning and nuanced personalization. However, efficiently managing and reasoning over long-term spatio-temporal memory remains a key challenge.

Parametric Memory Retrieval. While recent knowledge editing methods (Fang et al., 2024; Wang et al., 2024c) claim they can localize and modify specific representations, enabling models to selectively retrieve knowledge from their own parameters remains an open challenge. Efficient retrieval and integration of latent knowledge could significantly enhance memory utilization and reduce dependence on external indexing and memory management.

Lifelong Learning requires AI agents to continually integrate new information while retaining prior knowledge (Feng et al., 2024), necessitating robust memory systems to balance stability and plasticity. Parametric memory (Tian et al., 2024) enables in-weight knowledge adaptation but is vulnerable to forgetting, while structural memory (e.g., knowledge graph, tables) supports modular, targeted updates (Rasmussen et al., 2025). Unstructured memory, such as vector stores or raw dialogue

histories, offers flexible retrieval but requires dynamic compression and relevance filtering (Bae et al., 2022). Integrating these memory types under a continual learning framework—with mechanisms like consolidation, selective forgetting, and interleaved training—is essential for building adaptive, personalized lifelong agents capable of long-term memory management.

Brain-Inspired Memory Models. Memory in biological systems offers key insights for building more resilient and adaptive AI memory architectures. The brain manages the stability-plasticity dilemma through complementary learning systems: the hippocampus encodes fast-changing episodic experiences, while the cortex slowly integrates stable long-term memory (McClelland et al., 1995; Kumaran et al., 2016). Inspired by this, AI models increasingly adopt dual-memory architectures, synaptic consolidation, and experience replay to mitigate forgetting (Ritter et al., 2018; Wang et al., 2021). Cognitive concepts like memory reconsolidation (Dudai et al., 2015), bounded memory capacity (Cowan, 2001), and compartmentalized knowledge (Franklin et al., 2020) further inform strategies for update-aware recall, efficient storage, and context-sensitive generalization.

Meanwhile, the K-Line Theory (Minsky, 1980) points out that hierarchical memory structures are fundamental to biological cognition. These structures enable humans to efficiently organize memory across different levels of abstraction.—as seen in how infants group specific objects like "apple" and "banana" into broader categories like "fruit" and "food." Organizing the memory of AI systems with hierarchy structures for scalability and efficiency raises new challenges (Wang et al., 2024l; Han et al., 2025) and future directions (Wang et al., 2024k; Hong et al., 2024) for memory research.

Unified Memory Representation. While parametric memory (Yang et al., 2024) provides compact and implicit knowledge storage, and external memory (Zhong et al., 2024) offers explicit and interpretable information, unifying their representational spaces and establishing joint indexing mechanisms is essential for effective memory consolidation and retrieval. Future work should focus on developing unified memory representation frameworks that support shared indexing, hybrid storage, and memory operations across modalities and knowledge forms.

Multi-agent Memory. In multi-agent sys-

https://www.me.bot

⁴https://ima.qq.com

tems, memory is not only individual but also distributed—agents must manage their own internal memories while interacting with and learning from others. This raises unique challenges such as memory sharing, alignment, conflict resolution, and consistency across agents. Effective multi-agent memory systems should support both local retention of personalized experiences and global coordination through shared memory spaces or communication protocols. Future work may explore decentralized memory architectures, cross-agent memory synchronization, and collective memory consolidation to enable collaborative planning, reasoning, and long-term coordination.

Memory Threats & Safety. While memory significantly enhances the utility of LLMs by enabling up-to-date and personalized responses, its management remains a critical safety concern. Memory often stores sensitive and confidential data, making operations like adding or removing information far from trivial. Recent research has exposed serious vulnerabilities in memory handling, particularly in machine unlearning techniques designed to selectively erase data. Multiple studies (Liu et al., 2025b; Barez et al., 2025) have demonstrated that these methods are prone to malicious attacks which strengthens the need for more secure and reliable memory operations.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J. Nair, Ilya Soloveychik, and Purushotham Kamath. 2024. Keyformer: Kv cache reduction through key tokens selection for efficient generative inference. In *Proceedings of Machine Learning and Systems*, volume 6, pages 114–127.
- Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.
- Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako

- Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP* 2022, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks.
- Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. 2025. Open problems in machine unlearning for ai safety.
- Vincent-Pierre Berges, Barlas Oğuz, Daniel Haziza, Wen tau Yih, Luke Zettlemoyer, and Gargi Ghosh. 2024. Memory layers at scale.
- Shuyang Cao and Lu Wang. 2024. AWESOME: GPU memory-constrained long document summarization using memory mechanism and global salient content. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5925–5941, Mexico City, Mexico. Association for Computational Linguistics.
- Zhiwei Cao, Qian Cao, Yu Lu, Ningxin Peng, Luyang Huang, Shanbo Cheng, and Jinsong Su. 2024. Retaining key information under high compression ratios: Query-guided compressor for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12685–12695, Bangkok, Thailand. Association for Computational Linguistics.
- Harrison Chase. 2022. Langchain. https://www.langchain.com. Accessed: 2025-04-17.
- Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Pro-*

- ceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12041–12052.
- Mingda Chen, Yang Li, Karthik Padthe, Rulin Shao, Alicia Sun, Luke Zettlemoyer, Gargi Ghosh, and Wentau Yih. 2024a. Improving factuality with explicit working memory. *arXiv preprint arXiv:2412.18069*.
- Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. 2024b. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. *arXiv preprint arXiv:2402.11975*.
- Wenhu Chen, Zhihao He, Yu Su, Yunyao Yu, William Wang, and Xifeng Yan. 2021. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yilong Chen, Guoxia Wang, Junyuan Shang, Shiyao Cui, Zhenyu Zhang, Tingwen Liu, Shuohuan Wang, Yu Sun, Dianhai Yu, and Hua Wu. 2024c. NACL: A general and effective KV cache eviction framework for LLM at inference time. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7913–7926, Bangkok, Thailand. Association for Computational Linguistics.
- Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. arXiv preprint arXiv:2405.13792.
- Tsz Ting Chung, Leyang Cui, Lemao Liu, Xinting Huang, Shuming Shi, and Dit-Yan Yeung. 2024. Selection-p: Self-supervised task-agnostic prompt compression for faithfulness and transferability. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11057–11070, Miami, Florida, USA. Association for Computational Linguistics.
- Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.
- Coze. 2024. Coze: Build your own ai agent. https://www.coze.cn/. Accessed: April 19, 2025.
- Bhavana Dalvi Mishra, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9465–9480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurélie C. Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jirí Navrátil, Soham Dan, and Pin-Yu Chen.

- 2024. Larimar: Large language models with episodic memory control. In *ICML*.
- Ronald L Davis and Yi Zhong. 2017. The biology of forgetting—a perspective. *Neuron*, 95(3):490–503.
- N De Cao, W Aziz, and I Titov. 2021. Editing factual knowledge in language models. In *EMNLP* 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, pages 6491–6506
- Xuanwen Ding, Jie Zhou, Liang Dou, Qin Chen, Yuanbin Wu, Arlene Chen, and Liang He. 2024. Boosting large language models with continual learning for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4367–4377, Miami, Florida, USA. Association for Computational Linguistics.
- Harry Dong, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, and Beidi Chen. 2024. Get more with LESS: Synthesizing recurrence with KV cache compression for efficient LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11437–11452. PMLR.
- Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- Xinya Du, Sha Li, and Heng Ji. 2022. Dynamic global memory for document-level argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5264–5275, Dublin, Ireland. Association for Computational Linguistics.
- Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. *arXiv* preprint *arXiv*:2402.16288.
- Zhihua Duan and Jialin Wang. 2024. Exploration of llm multi-agent application implementation based on langgraph+ crewai. *arXiv preprint arXiv:2411.18241*.

- Yadin Dudai, Avi Karni, and Jan Born. 2015. The consolidation and transformation of memory. *Neuron*, 88(1):20–32.
- Ritam Dutt, Kasturi Bhattacharjee, Rashmi Gangadharaiah, Dan Roth, and Carolyn Rose. 2022. Perkgqa: Question answering over personalized knowledge graphs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 253–268.
- Ronen Eldan and Mark Russinovich. 2024. Who's harry potter? approximate unlearning for LLMs.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representa*tions.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. arXiv preprint arXiv:2410.02355.
- Yujie Feng, Xu Chu, Yongxin Xu, Guangyuan Shi, Bo Liu, and Xiao-Ming Wu. 2024. TaSL: Continual dialog state tracking via task skill localization and consolidation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1266–1279, Bangkok, Thailand. Association for Computational Linguistics.
- Nicholas T Franklin, Kenneth A Norman, Charan Ranganath, Jeffrey M Zacks, and Samuel J Gershman. 2020. Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, 127(3):327–361.
- Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. 2022. There are a thousand hamlets in a thousand people's eyes: Enhancing knowledge-grounded dialogue with personal memory. arXiv preprint arXiv:2204.02624.
- Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. 2025. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2024. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

- Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. LM-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008, Mexico City, Mexico. Association for Computational Linguistics.
- Kaiqiao Han, Tianqing Fang, Zhaowei Wang, Yangqiu Song, and Mark Steedman. 2025. Concept-reversed Winograd schema challenge: Evaluating and improving robust reasoning in large language models via abstraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 229–243, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yongchang Hao, Mengyao Zhai, Hossein Hajimir-sadeghi, Sepidehsadat Hosseini, and Frederick Tung. 2025. Radar: Fast long-context decoding for any transformer. In *The Thirteenth International Conference on Learning Representations*.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. IN-SPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.
- Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024a. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.
- Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024b. Never lost in the middle: Mastering long-context question answering with positionagnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642, Bangkok, Thailand. Association for Computational Linguistics.
- Yang He, Ruijie Fang, Isil Dillig, and Yuepeng Wang. 2025. Graphiti: Bridging graph and relational database queries. *arXiv preprint arXiv:2504.03182*.
- Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt W. Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. 2024c. Human-inspired perspectives: A survey on ai long-term memory. arXiv preprint arXiv:2411.00489.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop

- qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.
- Ruixin Hong, Hongming Zhang, Xiaoman Pan, Dong Yu, and Changshui Zhang. 2024. Abstraction-ofthought makes language models better reasoners. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1993–2027, Miami, Florida, USA. Association for Computational Linguistics.
- Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In *Advances in Neural Information Processing Systems*, volume 37, pages 1270–1303. Curran Associates, Inc.
- Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7. ACM.
- Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. 2023. Groundnlq@ ego4d natural language queries challenge 2023. arXiv preprint arXiv:2306.15255.
- Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.
- Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023a. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540, Singapore. Association for Computational Linguistics.
- Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2024. Learning retrieval augmentation for personalized dialogue generation. *arXiv* preprint *arXiv*:2406.18847.
- Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, et al. 2023b. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.

- LangChain Inc. 2025. Langgraph: Build resilient language agents as graphs. https://github.com/langchain-ai/langgraph. Accessed: 2025-04-17.
- Kai Tzu iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung won Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv* preprint arXiv:2112.09118.
- Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*.
- Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. IterCQR: Iterative conversational query reformulation with retrieval guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8121–8138, Mexico City, Mexico. Association for Computational Linguistics.
- Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. Advances in Neural Information Processing Systems, 37:12581–12611.
- Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo Angel, and Sijia Liu. 2024a. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. In *Annual Conference on Neural Information Processing Systems*.
- Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024b. Soul: Unlocking the power of second-order optimization for llm unlearning. arXiv preprint arXiv:2404.18239.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024a. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.
- Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, Mengyue Wu, Weizhi Ma, Mengdi Wang, and Tianqiao Chen. 2024b. Long term memory: The foundation of ai self-evolution. *arXiv* preprint *arXiv*:2410.15665.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *The Thirteenth International Conference on Learning Representations*.
- Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. 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)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Christopher Kiley and Colleen M Parks. 2022. Mechanisms of memory updating: State dependency vs. reconsolidation. *Journal of cognition*, 5(1):7.
- Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. 2024a. Dialsim: A real-time simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv* preprint arXiv:2406.13144.
- Seo Hyun Kim, Keummin Ka, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. 2024b. Ever-evolving memory by blending and refining the past. *arXiv preprint arXiv:2403.04787*.
- kingjulio8238. 2025. Memary. https://github.
 com/kingjulio8238/Memary. Accessed:
 2025-04-17.

- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Dharshan Kumaran, Demis Hassabis, and James L Mc-Clelland. 2016. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Dongkyu Lee, Chandana Satya Prakash, Jack FitzGerald, and Jens Lehmann. 2024a. Matter: Memoryaugmented transformer using heterogeneous knowledge sources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16110–16121.
- Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024b. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26396–26415. PMLR.
- Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024c. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.
- Mingcong Lei, Yiming Zhao, Ge Wang, Zhixin Mai, Shuguang Cui, Yatong Han, and Jinke Ren. 2025. Stma: A spatio-temporal memory agent for long-horizon embodied task planning. *arXiv preprint arXiv:2502.10177*.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.
- Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. 2025. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. 2024b. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550.
- Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. 2024c. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786, Miami, Florida, USA. Association for Computational Linguistics.
- Xiaonan Li and Xipeng Qiu. 2023. Mot: Memory-ofthought enables chatgpt to self-improve. In *Proceed*ings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 6354–6374, Singapore. Association for Computational Linguistics
- Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *arXiv preprint arXiv:2305.13269*.
- Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024d. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11851–11861, Bangkok, Thailand. Association for Computational Linguistics.
- Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024e. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 22947–22970. Curran Associates, Inc.
- Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2024f. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *arXiv preprint arXiv:2408.03615*.
- Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024g. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.
- Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. 2024h. Prompt compression for large language models: A survey.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.
- Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024b. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266.
- Di Liu, Meng Chen, Baotong Lu, Huiqiang Jiang, Zhenhua Han, Qianxi Zhang, Qi Chen, Chengruidong Zhang, Bailu Ding, Kai Zhang, Chen Chen, Fan Yang, Yuqing Yang, and Lili Qiu. 2024c. Retrievalattention: Accelerating long-context llm inference via vector retrieval.
- Jerry Liu. 2022. Llamaindex. https://www.llamaindex.ai. Accessed: 2025-04-17.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025a. A survey of personalized large language models: Progress and future directions. *arXiv* preprint arXiv:2502.11528.
- Minqian Liu, Shiyu Chang, and Lifu Huang. 2022a. Incremental prompting: Episodic memory prompt for lifelong event detection. arXiv preprint arXiv:2204.07275.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024d. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023a. RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.
- Shuai Liu, Hyundong J Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023b. Recap: retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. *arXiv* preprint arXiv:2306.07206.
- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022b. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv preprint arXiv:2209.00179*.
- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023c. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proceedings of ICLR*.

- Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024e. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1817–1829.
- Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023d. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. 2024f. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32332–32344. PMLR.
- Ziyao Liu, Huanyi Ye, Chen Chen, Yongsen Zheng, and Kwok-Yan Lam. 2025b. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*, 6:413–425.
- Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv* preprint arXiv:2308.08239.
- Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3281, Bangkok, Thailand. Association for Computational Linguistics.
- Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. UniTranSeR: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 103–114, Dublin, Ireland. Association for Computational Linguistics.
- Aru Maekawa, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2023. Generative replay inspired by hippocampal memory indexing for continual language learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 930–942.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.
- Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. 2022. Dsi++: Updating transformer memory with new documents. arXiv preprint arXiv:2212.09744.
- Daniel Mela, Aitor González-Agirre, Javier Hernando, and Marta Villegas. 2024. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5831–5847.
- memodb io. 2025. Memobase: Profile-based long-term memory for ai applications. https://github.com/memodb-io/memobase. Accessed: 2025-04-26.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Massediting memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Massediting memory in a transformer. In *The Eleventh International Conference on Learning Representations*.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Repre*sentations.
- Marvin Minsky. 1980. K-lines: A theory of memory. *Cognitive Science*, 4(2):117–133.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022b. Fast model editing at scale. In *International Conference on Learning Representations*.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022c. Memorybased model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Neo4j. 2012. Neo4j the world's leading graph database. Accessed: 2025-04-25.
- Yuxiang Nie, Heyan Huang, Wei Wei, and Xian-Ling Mao. 2022. Capturing global structural information in long document question answering with compressive graph selector network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5036–5047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Cicero Nogueira dos Santos, James Lee-Thorp, Isaac Noble, Chung-Ching Chang, and David Uthus. 2024. Memory augmented language models through mixture of word experts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4425–4438, Mexico City, Mexico. Association for Computational Linguistics.
- Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2022. UniK-QA: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1535–1546, Seattle, United States. Association for Computational Linguistics.
- OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt.
- OpenAI. 2025. Openai platform documentation: Embeddings guide. https://platform.openai.com/docs/guides/embeddings. Accessed: 2025-04-17.
- Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

- Junyeong Park, Junmo Cho, and Sungjin Ahn. 2025. Mr.steve: Instruction-following agents in minecraft with what-where-when memory. In *International Conference on Learning Representations (ICLR)*. Accepted as a poster at ICLR 2025.
- Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.
- Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *International Conference on Machine Learning*, pages 40034–40050. PMLR.
- USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, et al. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.
- Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.
- Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. 2023. Agent-om: Leveraging llm agents for ontology matching. *arXiv preprint arXiv:2312.00326*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings* of the 62nd Annual Meeting of the Association for

- Computational Linguistics (Volume 1: Long Papers), pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.
- Steven Ritter, Jane X Wang, Zeb Kurth-Nelson, Siddhant Jayakumar, Charles Blundell, and Timothy Lillicrap. 2018. Meta-learning through hebbian plasticity in random networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.
- Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. Nist Special Publication Sp, 109:109.
- Ali Safaya and Deniz Yuret. 2024. Neurocache: Efficient vector retrieval for long-range language modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 870–883, Mexico City, Mexico. Association for Computational Linguistics.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv* preprint *arXiv*:2304.11406.
- Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. 2024. Mastering memory tasks with world models. In *The Twelfth International Conference on Learning Representations*.
- Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. 2024. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985.
- Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: high-throughput generative inference of large language models with a single gpu. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.
- Haizhou Shi and Hao Wang. 2023. A unified approach to domain incremental learning with memory: Theory and algorithm. *Advances in Neural Information Processing Systems*, 36:15027–15059.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.
- Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. 2015. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766.

- Jianlin Su, Murtadha Ahmed, Bo Wen, Luo Ao, Mingren Zhu, and Yunfeng Liu. 2024. Naive Bayes-based context extension for large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7791–7807, Mexico City, Mexico. Association for Computational Linguistics.
- Xin Su, Tiep Le, Steven Bethard, and Phillip Howard. 2023. Semi-structured chain-of-thought: Integrating multiple sources of knowledge for improved language model reasoning. *arXiv preprint arXiv:2311.08505*.
- Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8211–8227, Miami, Florida, USA. Association for Computational Linguistics.
- Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 641–651. Association for Computational Linguistics.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024a. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv e-prints*, pages arXiv–2401.
- Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024b. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6459–6475, Miami, Florida, USA. Association for Computational Linguistics.
- Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2021. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *arXiv preprint arXiv:2112.09737*.
- Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. 2024. QUEST: Queryaware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International*

- Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 47901–47911. PMLR.
- Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023a. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv* preprint arXiv:2305.11482.
- Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023b. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468, Toronto, Canada. Association for Computational Linguistics.
- Deshraj Yadav Taranjeet Singh. 2024. Mem0: The memory layer for your ai agents. https://github.com/mem0ai/mem0.
- Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*.
- Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1524–1537.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. A MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2024a. Enhancing large language model with self-controlled memory framework.
- Bingbing Wang, Yiming Du, Bin Liang, Zhixin Bai, Min Yang, Baojun Wang, Kam-Fai Wong, and Ruifeng Xu. 2025a. A new formula for sticker retrieval: Reply with stickers in multi-modal and multi-session conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25327–25335.

- Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. 2021. Dualsystem episodic control: Integrating episodic memory and reinforcement learning. *Nature Human Behaviour*, 5(3):293–307.
- Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024b. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. 2022. Memory replay with data compression for continual learning. *arXiv preprint arXiv:*2202.06592.
- Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024c. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797.
- Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024d. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.
- Qingyue Wang, Yanan Fu, Yanan Cao, Shi Wang, Zhiliang Tian, and Liang Ding. 2025b. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, page 130193.
- Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025c. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193.
- Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. 2024e. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? arXiv preprint arXiv:2410.15267.
- Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024f. Symbolic working memory enhances language models for complex rule application. *arXiv* preprint arXiv:2408.13654.
- Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024g. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Shah, Yujia Bao, Yang Liu,

- and Wei Wei. 2025d. LLM unlearning via loss adjustment with only forget data. In *The Thirteenth International Conference on Learning Representations*.
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. 2024h. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143*.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023a. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.
- Ying Wang, Yanlai Yang, and Mengye Ren. 2023b. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*.
- Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. 2024i. Memoryllm: towards self-updatable large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 50453–50466.
- Yu Wang, Chi Han, Tongtong Wu, Xiaoxin He, Wangchunshu Zhou, Nafis Sadeq, Xiusi Chen, Zexue He, Wei Wang, Gholamreza Haffari, et al. 2024j. Towards lifespan cognitive systems. arXiv preprint arXiv:2409.13265.
- 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*.
- Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025e. Delta: An online document-level translation agent based on multi-level memory. In *International Conference on Learning Representations (ICLR)*.
- Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Wong, and Simon See. 2024k. AbsInstruct: Eliciting abstraction ability from LLMs through explanation tuning with plausibility estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–994, Bangkok, Thailand. Association for Computational Linguistics.
- Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024l. AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010, Mexico City, Mexico. Association for Computational Linguistics.

- Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024m. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4891–4906, Miami, Florida, USA. Association for Computational Linguistics.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024a. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.
- Wei Wu, Zhuoshi Pan, Chao Wang, Liyi Chen, Yunchu Bai, Tianfu Wang, Kun Fu, Zheng Wang, and Hui Xiong. 2025. Tokenselect: Efficient long-context inference and length extrapolation for llms via dynamic token-level kv cache selection.
- Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886.
- Yichen Wu, Hong Wang, Peilin Zhao, Yefeng Zheng, Ying Wei, and Long-Kai Huang. 2024b. Mitigating catastrophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization. In *Forty-first International Conference on Machine Learning*.
- Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2022. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5184–5196, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- xAI. 2023. Grok. https://grok.com. Accessed: 2025-04-19.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RE-COMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.
- Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for llms: A survey. *arXiv* preprint arXiv:2403.08319.

- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.
- Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*.
- Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Kang Liu, and Jun Zhao. 2024c. Generate-on-graph: Treat llm as both agent and kg in incomplete knowledge graph question answering. *arXiv preprint arXiv:2404.14741*.
- Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multi-Instruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada. Association for Computational Linguistics.
- Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, et al. 2024. *memory*³: Language modeling with explicit memory. *arXiv* preprint arXiv:2407.01178.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475.
- Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2029. Association for Computational Linguistics.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Haofei Yu, Cunxiang Wang, Yue Zhang, and Wei Bi. 2023. TRAMS: Training-free memory selection for long-range language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4966–4972, Singapore. Association for Computational Linguistics.
- Xihang Yue, Linchao Zhu, and Yi Yang. 2024. FragRel: Exploiting fragment-level relations in the external memory of large language models. In *Findings of*

- the Association for Computational Linguistics: ACL 2024, pages 16348–16361, Bangkok, Thailand. Association for Computational Linguistics.
- Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. 2023a. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.
- Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. 2023b. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.
- Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2023c. Llm-based medical assistant personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*.
- Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024a. LLM-based medical assistant personalization with short- and long-term memory coordination. In *Proceedings of the 2024 Conference* of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2386–2398, Mexico City, Mexico. Association for Computational Linguistics.
- Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024b. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.
- Taolin Zhang, Qizhou Chen, Dongyang Li, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, et al. 2024c. Dafnet: Dynamic auxiliary fusion for sequential model editing in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1588–1602.
- Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024d. A survey on the memory mechanism of large language model based agents. *arXiv* preprint arXiv:2404.13501.
- Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023d. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Wenting Zhao, Ye Liu, Tong Niu, Yao Wan, Philip Yu, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. 2024a. DIVKNOWQA: Assessing the reasoning ability of LLMs via open-domain question answering over knowledge base and text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 51–68, Mexico City, Mexico. Association for Computational Linguistics.
- Wenting Zhao, Ye Liu, Tong Niu, Yao Wan, Philip S. Yu, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. 2024b. DIVKNOWQA: Assessing the reasoning

ability of LLMs via open-domain question answering over knowledge base and text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 51–68. Association for Computational Linguistics.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. 2024c. Atom: Lowbit quantization for efficient and accurate llm serving. In *MLSys*.

Zhengyi Zhao, Shubo Zhang, Yiming Du, Bin Liang, Baojun Wang, Zhongyang Li, Binyang Li, and Kam-Fai Wong. 2025. Eventweave: A dynamic framework for capturing core and supporting events in dialogue systems. *arXiv preprint arXiv:2503.23078*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200.

Yun Zhu, Jia-Chen Gu, Caitlin Sikora, Ho Ko, Yinxiao Liu, Chu-Cheng Lin, Lei Shu, Liangchen Luo, Lei Meng, Bang Liu, and Jindong Chen. 2025. Accelerating inference of retrieval-augmented generation via sparse context selection. In *The Thirteenth International Conference on Learning Representations*.

Appendix

A Relative Citation Index

In this work, we identify impactful works by Relative Citation Index (RCI) metirc, which take the

publication age into account to prevent bias between original citations from different publication dates. The age A_i of a paper p_i is computed as:

$$A = T - Year_i + 1 \tag{1}$$

, where T is the date when the citation is collected (20th April 2025) and $Year_i$ is the year where paper i is first published. Thus, we can model the relation between citation number C_i and age A_i of paper p_i as:

$$\log(C_i + 1) = \beta + \alpha \log A_i + \epsilon_i \tag{2}$$

We collect papers from past 3 years (2022 to 2025) from Top NLP and ML conferences (i.e., ACL, NAACL, EMNLP, NeurIPS, ICML, ICLR). To reduce the bias from different research area, we use GPT to score the relevance of a paper with the four challenges discussed in the paper. We pick all the papers with score equal and higher than 8 and collect their publication date and citation numbers from Semantic Scholar API⁵. For papers without publication date field, we use the first conference day as the publication date. We gather a total number of 3,932 valid papers after the processing and compute the estimated $b\hat{e}ta$ and $\hat{\alpha}$ accordingly⁶. After that⁷, we are able to obtain the expected citation number \hat{C}_i of paper p_i with age A_i as:

$$\hat{C}_i = \exp(\hat{\beta}) + A_i^{\hat{\alpha}} \tag{3}$$

Then we compute the relative citation index RCI_i of paper p_i as:

$$RCI_i = \frac{C_i}{\hat{C}_i} \tag{4}$$

When $RCI_i >= 1$, we consider this paper overcited than its expectations, and vice versa. In this paper, we focus on the paper with RCI >= 1, for which we believe has more influence.

⁵https://www.semanticscholar.org/ product/api

⁶Noted that not all papers mentioned in this work are considered in estimating $\hat{\beta}$ and $\hat{\alpha}$, but they will be assigned a RCI score based on the publication age.

⁷The estimation is: $\hat{\beta} = 1.878$, $\hat{\alpha} = 1.297$

Datasets	Мо	Operations	DS Type	Per	TR	Metrics	Purpose	Year	Access
LongMemEval (Wu et al., 2024a)	text	Indexing, Retrieval, Compression	MS	×	/	Recall@K, NDCG@K, Accuracy	Benchmark chat assistants on long-term memory abilities, including temporal reasoning.	2025	[LINK]
LoCoMo (Maharana et al., 2024)	text + image	Indexing, Retrieval, Compression	MS	×	1	Accuracy, ROUGE, Preci- sion, Recall, F1	Evaluate long-term memory in LLMs across QA, event summarization, and multimodal dialogue tasks.	2024	[LINK]
MemoryBank (Zhong et al., 2024)	text	Updating, Retrieval	MS	1	X	Accuracy, Hu- man Eval	Enhance LLMs with long-term memory capabilities, adapting to user personalities and contexts.	2024	[LINK]
PerLTQA (Du et al., 2024)	text	Retrieval	MS	•	х	MAP, Recall, Precision, F1, Accuracy, GPT4 score	To explore personal long-term memory question answering ability.	2024	[LINK]
MALP (Zhang et al., 2024a)	text	Retrieval, Compression	QA	✓	×	ROUGE, Accuracy, Win Rate	Preference-conditioned dialogue generation. Parameter-efficient fine-tuning (PEFT) for customization.	2024	[LINK]
DialSim (Kim et al., 2024a)	text	Retrieval	MS	1	X	Accuracy	To evaluate dialogue systems under realistic, real-time, and long-context multiparty conversation conditions.	2024	[LINK]
CC (Jang et al., 2023)	text	Retrieval	MS	X	1	BLEU, ROUGE	For long-term dialogue modeling with time and relationship context.	2023	[LINK]
LAMP (Salemi et al., 2023)	text	Consolidation, Retrieval, Compression	MS	•	1	Accuracy, F1, ROUGE	Multiple entries per user. Supports both user-based splits and time-based splits, enabling evaluation of short-term and long-term personalization.	2023	[LINK]
MSC (Xu et al., 2021)	text	Consolidation, Retrieval, Compression	MS	•	×	PPL	To evaluate and improve long-term dia- logue models via multi-session human- human chats with evolving shared knowledge.	2022	[LINK]
DuLeMon (Xu et al., 2022)	text	Consolidation, Updating Retrieval, Compression	MS	•	×	Accuracy, F1, Recall, Pre- cision, PPL, BLEU, DIS- TINCT	For dynamic persona tracking and consistent long-term human-bot interaction.	2022	[LINK]
2WikiMultiHopQA (Ho et al., 2020)	table + knowl- edge base + text	Consolidation, Indexing, Retrieval, Compression	QA	×	×	EM, F1	Multi-hop QA combining structured and unstructured data with reasoning paths.	2020	[LINK]
NQ (Kwiatkowski et al., 2019)	text	Retrieval, Compression	QA	×	Х	EM, Fl	Open-domain QA based on real Google search queries.	2019	[LINK]
HotpotQA (Yang et al., 2018)	text	Retrieval, Compression	QA	×	×	EM, F1	Multi-hop QA with explainable reasoning and sentence-level supporting facts.	2018	[LINK]

Table 2: Datasets used for evaluating **long-term memory**. "Mo" denotes modality. "Ops" denotes operability (placeholder). "DS Type" indicates dataset type (QA – question answering, MS – multi-session dialogue). "Per" and "TR" indicate whether persona and temporal reasoning are present.

Datasets	Modality	Operations	Metrics	Purpose	Year	Access
WikiText-103 (Merity et al., 2017)	text	compression	PPL	Corpus with 100 million tokens extracted from the set of verified articles on Wikipedia for long context language modeling.	2016	[LINK]
PG-19 (Rae et al., 2020)	text	compression	PPL	Corpus constructed with books extracted from the Project Gutenberg books library for long context language modeling.	2019	[LINK]
LRA (Tay et al., 2021)	text + image	compression, retrieval	Acc	Benchmark constructed with 6 identical tasks for evaluating efficient long context language models.	2020	[LINK]
NarrativeQA (Kočiský et al., 2018)	text	retrieval	Bleu-1, Bleu-4, Meteor, Rouge-L, MRR	Question Answering dataset could be used for evaluating long context QA ability.	2017	[LINK]
TriviaQA (Joshi et al., 2017)	text	retrieval	EM, F1	Question Answering dataset could be used for evaluating long context QA ability.	2017	[LINK]
NaturalQuestions (Kwiatkowski et al., 2019)	text	retrieval	EM, F1	Question Answering dataset could be used for evaluating long context QA ability.	2019	[LINK]
MusiQue (Trivedi et al., 2022)	text	retrieval	F1	Challenging multi-hop Question Answering dataset for evaluating long context reasoning and QA ability.	2021	[LINK]
CNN/DailyMail (Nallapati et al., 2016)	text	compression	Rouge-1, Rouge-2, Rouge-L	Over 300k news articles from CNN and Dai- lyMail for evaluating long document sum- marization	2016	[LINK]
GovReport (Huang et al., 2021)	text	compression	Rouge-1, Rouge-2, Rouge-L, Bert Score	Reports written by government research agencies for evaluating long document sum- marization	2021	[LINK]
L-Eval (An et al., 2024)	text	compression, retrieval	Rouge-L, F1, GPT4	Benchmark containing 20 sub-tasks spe- cially designed for evaluating long context language models from different aspect.	2023	[LINK]
LongBench (Bai et al., 2024)	text	compression, retrieval	F1, Rouge-L, Accuracy, EM, Edit Sim	Benchmark containing 14 English tasks, 5 Chinese tasks, and 2 code tasks for systematical long context evaluation.	2023	[LINK]
LongBench v2 (Bai et al., 2025)	text + table + KG	compression, retrieval	Acc	Updated version of LongBench which is much longer and more challenging, with consistent multi-choice format for reliable evaluation	2024	[LINK]

Table 3: Datasets for **long-context memory** evaluation.

Dataset	Modality	Operations	Metrics	Purpose	Year	Access
KnowEdit (Zhang et al., 2024b)	text	updating	Edit Success, Portability, Locality, and Fluency	Consists of 6 datasets . Provide a comprehensive evaluation covering knowledge insertion , modification , and erasure .	2024	[LINK]
MQUAKE-CF (Zhong et al., 2023)	text	updating	Edit-wise Success Rate, Instance-wise Accuracy, Multi-hop Accuracy	To evaluate the propagation of counterfactual knowledge editing affects through multi-hop reasoning, extending up to 4 hops, where a single reasoning chain may contain multiple edits.	2023	[LINK]
MQUAKE-T (Zhong et al., 2023)	text	updating	Edit-wise Success Rate, Instance-wise Accuracy, Multi-hop Accuracy	To evaluate the propagation of temporal knowledge editing affects through multi-hop reasoning, extending up to 4 hops, with only one edit per reasoning chain.	2023	[LINK]
Counterfact (Meng et al., 2022a)	text	updating	Efficacy Score, Efficacy Magnitude, Paraphrase Scores, Paraphrase Magnitude, Neighborhood Score, Neighborhood Magnitude	To evaluate substantial and improbable factual changes over superficial edits, especially those previously deemed unlikely by a model.	2022	[LINK]
zsRE (De Cao et al., 2021)	text	updating	Success Rate, Retain Accuracy, Equivalence Accuracy, Performance Deterioration	One of the earliest dataset used to evaluate knowledge editing.	2021	[LINK]
MUSE (Shi et al., 2024)	text	forgetting	VerbMem, KnowMem, PrivLeak	A comprehensive machine unlearning evaluation benchmark that enumerates six diverse desirable properties for unlearned models.	2024	[LINK]
KnowUnDo (Tian et al., 2024)	text	forgetting	Unlearn Success, Retention Success, Perplexity, ROUGE-L	A benchmark containing copyrighted content and user privacy domains to evaluate if the unlearning process inadvertently erases essential knowledge.	2024	[LINK]
RWKU (Jin et al., 2024)	text	forgetting	ROUGE-L	To evaluate real-world knowledge unlearning under practical , corpus-free conditions using real-world targets and adversarial assessments.	2024	[LINK]
WMDP (Li et al., 2024b)	text	forgetting	QA accuracy	Serve as a proxy measurement of hazardous knowledge in biosecurity , cybersecurity , and chemical security .	2024	[LINK]
TOFU (Maini et al., 2024)	text	forgetting	Probability, ROUGE, Truth Ratio	A novel unlearning dataset with facts about 200 fictitious authors .	2024	[LINK]
ABSA (Ding et al., 2024)	text	Consolidation	F1	A dataset for aspect-based sentiment analysis to evaluate LLMs in continual learning settings.	2024	[LINK]
SGD (Rastogi et al., 2020)	text	Consolidation	JGA, FWT (Forward Transfer), BWT (Backward Transfer)	A multi-turn task-oriented dialogue dataset that supports evolving user intents.	2020	[LINK]
INSPIRED (Hayati et al., 2020)	text	Consolidation	JGA, FWT (Forward Transfer), BWT (Backward Transfer)	A multi-turn task-oriented dialogue dataset that supports evolving user intents.	2020	[LINK]
Natural Question (Kwiatkowski et al., 2019)	text	Consolidation	Indexing Accuracy, Hits@1	A multi-purpose dataset that offers indexed documents and supports continual learning across evolving document collections.	2019	[LINK]

Table 4: Datasets for parametric memory evaluation.

Datasets	Mo	Ops	Src#	Mod#	Task	Metrics	Purpose	Year	Access
MultiChat (Wang et al., 2025a)	text + image	Retrieval	2	2	Retrieval	Precision, mAP, GPT-4	Image-grounded sticker retrieval with cross-session image-text dialogue context.	2025	[LINK]
Context-conflicting (Tan et al., 2024a)	text	Compression	2	1	Conflict	DiffGR, EM, Similarity	Designed to evaluate a model's ability to handle conflicting evidence across sources.	2024	[LINK]
EgoSchema (Mangalam et al., 2023)	video + text	Retrieval, Compression	3	2	Fusion	Accuracy	Combines episodic video memory, so- cial schema, and conversation for long- term memory QA.	2023	[LINK]
Ego4D NLQ (Hou et al., 2023)	video + text	Retrieval, Compression	2	2	Fusion	Recall@K	Video QA task focusing on natural lan- guage queries over egocentric video with temporal memory.	2022	[LINK]
2WikiMultihopQA (Ho et al., 2020)	text	Indexing, Retrieval, Compression	2	1	Reasoninį	EM, F1	Multi-hop QA requiring reasoning across two Wikipedia passages with sentence-level supporting evidence.	2020	[LINK]
HybridQA (Chen et al., 2021)	text	Retrieval Compression	2	1	Reasoning	EM, F1	QA requiring reasoning across struc- tured tables and unstructured text.	2020	[LINK]
CommonsenseVQA (Talmor et al., 2019)	text + image	Retrieval Compression	2	2	Fusion	Accuracy	Commonsense question answering over visual scenes requiring visual-textual fusion.	2019	[LINK]
NaturalQuestions (Kwiatkowski et al., 2019)	text	Retrieval Compression	>1*	1	Conflict	EM, F1	Real-world QA over Google search snip- pets; often used as source for contradic- tion analysis.	2019	[LINK]
ComplexWebQuestions (Talmor and Berant, 2018)	text	Retrieval Compression	>1*	1	Reasoning	EM, F1	Compositional QA requiring multi-step reasoning across web snippets.	2018	[LINK]
HotpotQA (Yang et al., 2018)	text	Retrieval Compression	2	1	Conflict	EM, F1, Sup- porting Fact Ac- curacy	Multi-hop QA with paragraph-level source documents and sentence-level supporting facts.	2018	[LINK]
TriviaQA (Joshi et al., 2017)	text	Retrieval Compression	≥6	1	Conflict	EM, F1	QA over trivia-style questions with noisy web sources; useful for source dis- agreement analysis.	2017	[LINK]
WebQuestionsSP (Yih et al., 2016)	text	Indexing Retrieval Compression	>1*	1	Reasoning	F1, Accuracy	Enhanced version of WebQuestions with structured reasoning chains.	2016	[LINK]
Flickr30K (Young et al., 2014)	text + image	Retrieval Compression	2	2	Retrieval	Similarity	Image-caption pairs widely used for cross-modal retrieval and alignment tasks.	2014	[LINK]

Table 5: Datasets used for evaluating **multi-source memory**. "Mo" denotes data modality. "Ops" indicates operations. "Src#" = number of information sources per instance; "Mod#" = number of modalities; "Task" = retrieval, fusion, reasoning, or conflict resolution.

Method	Type	TF	RE	Input	Output	LMs	Ops	Features	Year	Code
PERKGQA (Dutt et al., 2022)	Augmentation	1	•	Retrieved & Knowledge Graph + Query	Response	RoBERTa	Retrieval	long-term dialogue modeling, event & persona memory, mudular agent architecture	2022	[LINK]
CLV (Tang et al., 2023b)	Adaption	×	×	Persona + Query	Response	GPT-2	Consolidation	contrastive learning, clustered dense persona, dialogue generation	2023	[LINK]
RECAP (Liu et al., 2023a)	Augmentation	×	1	Retrieved & Context + Query	Response	Transformers	Retrieval	hierarchical transformer retriever, context-aware prefix encoder	2023	[LINK]
SiliconFriend (Zhong et al., 2024)	Augmentation	×	1	Retrieved & Context + Query	Response	ChatGLM-6B, BELLE-7B, gpt-3.5-turbo	Consolidation, Updating, Forgetting, Retrieval	fine-tuning, RAG, Ebbinghaus Forgetting	2024	[LINK]
MALP (Zhang et al., 2024a)	Adaption	×	1	Retrieved & Context + Query	Response	GPT3.5, LLaMA-7B, LLaMA-13B	Consolidation, Retrieval	memory coordination, computational bionic memory mechanism, patient profile, self-chat	2024	[LINK]
PERPCS (Tan et al., 2024b)	Adaption	×	×	User History	/	Llama-2-7B	Consolidation	modular PEFT sharing, collaborative personalization, user history assembly	2024	[LINK]
LAPDOG (Huang et al., 2023a)	Augmentation	1	1	Retrieved & Context + Query	Response	T5	Consolidation, Updating, Retrieval	Story-based persona retrieval, joint retriever-generator training	2024	[LINK]
LD-Agent (Li et al., 2024a)	Augmentation	1	1	Retrieved & Context + Query	Response	ChatGLM, BlenderBot, ChatGPT	Consolidation, Updating, Retrieval	long-term dialogue modeling, event & persona memory, mudular agent architecture	2025	[LINK]

Table 6: Overview of methods for **long-term memory in personalization**. "TF" (Training Free) denotes whether the method operates without additional gradient-based updates. "RE" (Retrieval Module) denotes whether the method needs Retrieval.

Method	Type	TF	RE	DS	Input	Output	LMs	Ops	Features	Year	Code
MemoChat (Lu et al., 2023)	Consolidation	×	/	,	Dialogue History + Query	Response	GPT4, ChatGPT, VIcuna-7B, 13B, 33B, T5	Consolidation, Retrieval	Structured memos, memory-driven dialogue, mem- orization-retrieval-response cycle	2023	[LINK]
MemoryBank (Zhong et al., 2024)	Consolidation	×	1	•	Retrieved & Context + Query	Response	ChatGLM-6B, BELLE-7B, gpt-3.5-turbo	Consolidation, Updating, Forgetting, Retrieval	fine-tuning, RAG, Ebbinghaus Forgetting	2024	[LINK]
NLI-Transfer (Bae et al., 2022)	Updating	/	1	1	Memory + Dialogue History	Response	T5	Consolidation, Updating, Retrieval	Session-level memory tracking, evolving dialogue system	2022	[LINK]
FLOW-RAG (Wang et al., 2024e)	Updating	×	1	×	Knowledge Base + Query	Response	GPT4o, Gemini, llama2-7B-chat	forgetting	RAG-based unlearning	2024	[LINK]
FLARE (Jiang et al., 2023b)	Retrieval	×	1	×	Database + Query	Response	WebGPT, WebCPM	retrieval	Active retrieval during generation, forward-looking query prediction	2023	[LINK]
HippoRAG (Gutiérrez et al., 2024)	Retrieval	×	1	×	Context + Query	Response	ColBERTv2, GPT-3.5-turbo, Llama-3.1-8B, 70B	Indexing	Hippocampal-inspired retrieval, multi-hop QA, Knowledge graph integration	2024	[LINK]
IterCQR (Jang et al., 2024)	Retrieval	×	1	1	Dialogue History + Query	Retrieved Results	Transformer++	Retrieval	Iterative query reformulation, context-aware query rewriting	2024	[LINK]
Chen et al., 2024a)	Memory Grounded Generation	/	/	×	Context	Response	Llama-3.1-70B, 8B	Updating, Retrieval	Explicit working memory, online fact-checking feedback, factual long-form generation	2025	[LINK]
MEMORAG (Qian et al., 2024)	Memory Grounded Generation	×	1	×	Context + Query	Response	Mistral7B-Instruct, Phi-3-mini-128K- instruct, GPT-40	Retrieval, Compression	Global memory retrieval, KV memory compression, Feedback-guided generation	2024	[LINK]
ReadAgent (Lee et al., 2024c)	Generation	×	1	×	Context + Query	Retrieved Passages/- Summary	PaLM 2	Updating, Retrieval	Episodic gist memory, dynamic memory retrieval, extended context window	2024	[LINK]
ICAL (Sarch et al., 2024)	Generation	×	x	×	Examples + Task Instruction	Trajectory + Thoughts	GPT4V, Qwen2VL	Updating	Trajectory abstraction memory, multi-modal, iterative reasoning correction	2025	[LINK]

Table 7: Overview of methods for **long-term memory in memory management and utilization**. "TF" (Training Free) denotes whether the method operates without additional gradient-based updates. "RE" (Retrieval Module) denotes whether the method needs Retrieval. "DS" (Dialogue System) denotes whether the method aims for a dialogue task.

Method	Туре	TF	DF	Operations	LMs	Features	Year	Code
StreamingLLM (Xiao et al., 2024)	KV Cache Dropping	1	×	Compression	Llama-2, MPT, PyThia, Falcon	Static KV cache dropping, Attention sink in the initial tokens	2023	[LINK]
FastGen (Ge et al., 2024)	KV Cache Dropping	1	×	Compression	Llama-1 7B/13B/30B/65B	Adaptive profiling-based KV cache dropping	2023	[LINK]
H ₂ O (Zhang et al., 2023d)	KV Cache Dropping	1	×	Compression	OPT, Llama-1, GPT-NeoX	Dynamica KV cache dropping, Retain Heavy Hitter tokens	2023	[LINK]
SnapKV (Li et al., 2024e)	KV Cache Dropping	1	×	Compression	LWM-Text-Chat-1M, LongChat-7b-v1.5-32k, Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1	Head-wise KV cache dropping, Attention head behavior	2024	[LINK]
LESS (Dong et al., 2024)	KV Cache Storing Optimization	×	/	Compression	Llama-2 13B, Falcon 7B	Low-rank KV cache storage, enable querying all tokens	2024	[LINK]
KIVI (Liu et al., 2024f)	KV Cache Storing Optimization	1	1	Compression	Llama-2 7B/13B, Llama-3 8B, Falcon 7B, Mistral-7B	Asymmetrical KV cache quantization	2024	[LINK]
KVQuant (Hooper et al., 2024)	KV Cache Storing Optimization	1	1	Compression	LLaMA-7B/13B/30B/65B, Llama-2-7B/13B/70B, Llama-3-8B/70B, and Mistral-7B	KV cache quantization	2024	[LINK]
QUEST (Tang et al., 2024)	KV Cache Selection	1	1	Retrieval	LongChat-7B-v1.5-32K, Yarn-Llama2-7B-128K	Query-aware KV cache selection	2024	[LINK]
TokenSelect (Wu et al., 2025)	KV Cache Selection	1	1	Retrieval	Qwen2 7B, Llama-3 8B, Yi-1.5-6B	Dynamic token-level KV cache selection	2024	[LINK]

Table 8: Overview of methods for **long-context memory in Parametric Efficiency**. "TF" (Training Free) denotes whether the method operates without additional gradient-based updates. "DF" (Dropping Free) denotes whether the method able to maintain all the KV cache without dropping.

Method	Туре	SM	TM	Operations	LMs	Features	Year	Code
GraphReader (Li et al., 2024c)	Context Selection	T	G	Retrieval	GPT-4-128k	Graph-based agent, Structuring long context to a graph	2024	[LINK]
Sparse RAG (Zhu et al., 2025)	Context Selection	T	P	Retrieval	Gemini	Sparse context selection, Reduce involved documents in decoding	2024	N/A
Ziya-Reader (He et al., 2024b)	Context Selection	T	T	Retrieval	Ziya2-13B-Base	Supervised finetuning, Position agnostic multi-step QA	2023	[LINK]
xRAG (Cheng et al., 2024)	Context Compression	T	P	Compression	Mistral-7b and Mixtral-8x7b	Soft prompt compression	2024	[LINK]
RECOMP (Xu et al., 2024a)	Context Compression	T	T	Compression	GPT-2, GPT2-XL, GPT-J, Flan-UL2	Hard prompt compression, extractive compressor, abstractive compressor	2023	[LINK]
LongLLMLingua (Jiang et al., 2024a)	Context Compression	T	T	Compression	GPT-3.5-Turbo-06136, LongChat-13B-16k	Hard prompt compression	2023	[LINK]
QGC (Cao et al., 2024)	Context Compression	T	T	Compression	LongChat-13B16K, LLaMA-2-7B	Query-guided dynamic context compression	2024	[LINK]

Table 9: Overview of methods for **long-context memory in Contextual Utilization**. "SM" (Source Modal) denotes the source modality of contextual memory. "TM" (Target Modal) denotes target modality (processed for selection / after compression) of contextual memory (T - Text, G - Graphs, P - Parametric).

Method	Type	PR	TF	BES	SEO	LMs	Main Advancement	Year	Code
AlphaEdit (Fang et al., 2025)	locating-then- editing	×	1	1	/	gpt2-xl-1.5b, gpt-j-6b, llama3-8b	Protect the preserved knowledge by projecting perturbation onto the null space. Add a regularization term when optimizing v* for sequential editing.	2024	[LINK]
MEMAT (Mela et al., 2024)	locating-then- editing	×	1	1	×	aguila-7b	MEMAT is expanded upon MEMIT with attention heads corrections for cross-lingual editing.	2024	[LINK]
MEMIT (Meng et al., 2023)	locating-then- editing	×	1	1	×	gpt-j-6b, gpt-neox-20b	Optimize a relaxed least-squares objective, enabling a simple closed-form solution for efficient massive batch editing.	2022	[LINK]
ROME (Meng et al., 2022a)	locating-then- editing	×	1	×	×	gpt2-x1-1.5b	The most classic locate-the-edit method. Perform a rank-one update on the weights of a single MLP layer.	2022	[LINK]
DAFNET (Zhang et al., 2024c)	meta learning	×	x	×	1	gpt-j-6b, llama2-7b	Supports sequential editing through Intra-editing Attention Flow (within facts) and Inter-editing Attention Flow (across facts).	2024	[LINK]
MEND (Mitchell et al., 2022b)	meta learning	×	x	1	×	gpt-neo gpt-j-6b t5-xl t5-xxl bert-base bart-base	More scalable and fast than KE. Decompose gradient into rank-one outer product form.	2021	[LINK]
KE (De Cao et al., 2021)	meta learning	×	×	1	×	bert-base, bart-base	The first one employs a hypernetwork to learn how to modify the gradient. Pose LSTM to project the sentence embedding into rank-1 mask over the gradient.	2021	[LINK]
IKE (Zheng et al., 2023)	prompt	/	1	-	-	gpt-j-6b, gpt2-xl-1.5b, gpt-neo, gpt-neox, opt-175b	The first use ICL to edit knowledge in LLMs.	2023	[LINK]
MeLLo (Zhong et al., 2023)	prompt	1	1	-	-	vicuna-7b, gpt-j-6b	Question Decompose + Self Check	2023	[LINK]
Larimar (Das et al., 2024)	additional parameters	1	1	1	1	gpt2-x1, gpt-j-6b	Introduce a decoupled latent memory module that conditions the LLM decoder at test time without parameter updates.	2024	[LINK]
MEMORYLLM (Wang et al., 2024i)	additional parameters	1	×	1	1	llama2-7b	Introduces a fixed-size memory pool in a frozen LLM that is incrementally and selectively updated with new knowledge.	2024	[LINK]
WISE (Wang et al., 2024c)	additional parameters	1	x	1	1	llama2-7b, mistral-7b, gpt-j-6b	Support sequential editing by Side Memory Design and Knowledge Sharding and Merging.	2024	[LINK]
CaliNET (Dong et al., 2022)	additional parameters	1	ж	1	×	t5-base, t5-large	Add the output of FFN-like CaliNET to the original FFN output.	2022	[LINK]
SERAC (Mitchell et al., 2022c)	additional parameters	1	×	1	1	t5-large, bert-base, blenderbot-90m	Scope Classifier + Counterfactual Model. Sequentially or simultaneously applying k edits yields the same edited model.	2022	[LINK]

Table 10: Overview of methods for **parametric memory optimization in editing**. "PR" (Parametric Reserving) indicates whether the method avoids direct modification of the model's internal weights. "TF" (Training-Free) denotes whether the method operates without traditional iterative optimization. "BES" (Batch Editing Support) reflects the method's ability to handle multiple edits simultaneously. "SEO" (Sequential Editing Optimization) specifies whether the method introduces mechanisms tailored for sequential Editing. "LMs" lists the language models used for empirical evaluation.

Method	Туре	PR	TF	BUS	SUO	LMs	Main Advancement	Year	Code
ULD (Ji et al., 2024)	additional parameters	1	×	1	×	llama2-chat-7b, mistral-7b-instruct	Derive the unlearned LLM by computing the logit difference between the target and the assistant LLMs.	2024	[LINK]
EUL (Chen and Yang, 2023)	additional parameters	1	×	1	/	t5-base, t5-3b	Introduce unlearning layers which are learned to forget requested data. Support sequential unlearning by using a fusion mechanism to merge different unlearning layers.	2023	[LINK]
ECO (Liu et al., 2024b)	prompt	1	x	1	×	68 llms ranging from 0.5b to 236b	ECO unlearns by corrupting prompt embeddings based on classifier detection without changing the model.	2024	[LINK]
ICUL (Pawelczyk et al., 2024)	prompt	1	1	-	-	bloom-560m, bloom-1.1b, bloom-3b, llama2-7b	The first use ICL for unlearning in LMs.	2023	[LINK]
WAGLE (Jia et al., 2024a)	locating-then- unlearning	×	×	1	×	llama2-7b-chat, zephyr-7b-beta, llama2-7b	WAGLE uses bi-level optimization to compute weight attribution scores that guide selective fine-tuning for efficient and modular unlearning.	2024	[LINK]

Table 11: Overview of methods for **parametric memory optimization in unlearning**. "PR" (Parametric Reserving) indicates whether the method avoids direct modification of the model's internal weights. "TF" (Training-Free) denotes whether the method operates without traditional iterative optimization. "BUS" (Batch Unlearning Support) reflects the method's ability to handle multiple edits simultaneously. "SUO" (Sequential Unlearning Optimization) specifies whether the method introduces mechanisms tailored for sequential Editing. "LMs" lists the language models used for empirical evaluation.

Method	Type	TF	TB	TS	Domain	LMs	Main Advancement	Year	Code
SELF- PARAM (Wang et al.)	Regularization- based Learning	×	×	Task- Free	Question Answering	T5	Employs a training objective that minimizes the Kullback-Leibler (KL) divergence between the predictions of the original model and target model.	2025	[LINK]
MBPA++ (Wang et al., 2024j)	Replay-based	×	×	CIL	None	REPLAY, MBPA	Integrate Maintaining a small, randomly selected subset (as low as 1%) of past examples in memory can achieve performance comparable to larger memory sizes.	2025	[LINK]
LSCS (Wang et al., 2024j)	Interactive Learning	×	×	CIL	Abstracting/ Merging/ Retrieval	/	Integrate multiple storage mechanisms and achieve both abstraction and experience merging and long-term retention with accurate recall.	2025	[LINK]
TaSL (Feng et al., 2024)	Regularization- based Learning	×	×	TIL	Dialogue System	T5, Llama-7B	Parameter-level task skill localization and consolidation enable knowledge transfer without memory replay.	2024	[LINK]
EMP (Liu et al., 2022a)	Replay-based	x	×	CLI	Event detection	BERT-ED, KCN	Design continuous prompts associated with each event type.	2023	[LINK]
UDIL (Shi and Wang, 2023)	Interactive Learning	×	/	DLI	Event detection	oEWC, SI, LwF, A-GEM, CLS-ER, ESM, etc.	Introducing adaptive coefficients that are optimized during training to achieve tighter generalization error bounds and better performance across domains.	2023	[LINK]
DSI++ (Mehta et al., 2022)	Replay-based	×	1	TIL	Information Retrieval	T5	Enables continual document indexing while retaining query performance on old and new data.	2022	[LINK]
MRDC (Wang et al., 2022)	Replay-based	×	1	CIL	Object detection	LUCIR, PODNet	Enhances memory replay by compressing data, balancing sample quality and quantity for continual learning.	2022	[LINK]

Table 12: Overview of methods for **parametric memory modification in continual learning**. "TB" denotes the task boundary whether exists. "TS" denotes the task settings including TIL (Task Incremental Learning), CIL (Class Incremental Learning), DIL (Domain Incremental Learning), Task-Free.

Method	Туре	TF	STs	SNs	Input	Output	LMs	Ops	Features	Year	Code
GoG (Xu et al., 2024c)	reasoning	1	KG + text	WebQSP, CWQ	KG + prompt + query	answer	GPT- 3.5,GPT-4, Qwen-1.5- 72B-Chat, LLaMA3- 70B- Instruct	Retrieval, Compression	integrate internal and external knowledge	2024	[LINK]
RKC-LLM (Wang et al., 2023a)	conflict	1	model + text	prompt + context	entities	answer	ChatGPT	Compression	Conflict span localization, instruction-guided conflict handling	2024	[LINK]
BGC-KC (Tan et al., 2024a)	conflict	1	model + text	AIG, AIR	documents + query	answer	GPT-4, GPT-3.5, Llama2- 13b, Llama2-7b	Retrieval, Compression	attribution tracing framework, evaluate LLM bias	2024	[LINK]
Sem-CoT (Su et al., 2023)	reasoning	×	Knowledge Graph + text +Model	Wikidata, 2Wiki, MuSiQue, TKB	CoT prompt + Query	answer	llama2-7b, 13b, 70b, 65b	Retrieval, Compression	Semi-structured prompting for multi-source input fusion	2023	[LINK]
CoK (Li et al., 2023)	reasoning	×	Database + Tables + Text	Wikidata, Wikipedia,and Wikitables, Flashcard, UpToDate, ScienceQA, CK-12	CoT prompt + Query	answer	gpt-3.5- turbo	Retrieval, Compression	Heterogeneous knowledge integration, dynamic knowledge retrieval, adaptive query generation across formats	2023	[LINK]
DIVKNOW((Zhao et al., 2024a)	reasoning	×	Knowledge Base + text	Wikidata, DIVKNOWQA	CoT prompt + Query	answer	gpt-3.5- turbo	Retrieval, Compression	Two-hop reasoning, symbolic query generation for structured data	2023	[LINK]
StructRAG (Li et al., 2024g)	reasoning	×	KG + Table + text	Loong, Podcast Transcripts	documents + query	answer	Qwen2-7B, 72B	Retrieval, Compression	Cognitive-inspired structurization, dynamic structure selection	2023	[LINK]

Table 13: Overview of methods for **multi-source memory in cross-textual integration**. "TF" (Training Free) denotes whether the method operates without additional gradient-based updates. "STs" denotes the source types. "SNs" denotes the source dataset names.

Method	Type	TF	DS	Mo	Input	Output	Modeling	Ops	Features	Year	Code
IGSR (Wang et al., 2025a)	retrieval	1	1	text + image	image- text dialogue	stickers	LLaVa, GPT4, Qwen-VL, CLIP, Llama3	retrieval	multi-modal memory bank, sticker retrieval, intention aware cross-session dialogue	2025	[LINK]
VISTA (Zhou et al., 2024)	retrieval	1	×	text + image	image- text query	retrieved response	CLIP, BLIP-B, Pic2Word	retrieval	Visual Token Injection, composed data fine-tuning	2024	[LINK]
UniVL-DR (Liu et al., 2022b)	retrieval	×	×	text + image	image- text query	retrieved response	VinVLDPR, CLIP-DPR	retrieval	Modality-balanced hard negatives	2023	[LINK]
MultiInstruct* (Xu et al., 2023)	fusion	1	×	text + image	instruction + instances	response	OFA	compression	Cross-modal transfer learning	2023	[LINK]
NextChat (Zhang et al., 2023b)	fusion	×	1	text + image + boxes	image + text	response	CLIP	compression	Cross-modal alignment	2023	[LINK]
UniTranSeR (Ma et al., 2022)	fusion	×	1	text + image	context	response	MLM + MPM	compression	Intention-aware response generation, unified transformer space	2022	[LINK]

Table 14: Overview of methods for **multi-source memory in Multi-modal Coordination**. "TF" (Training Free) denotes whether the method operates without additional gradient-based updates. "DS" (Dialogue System) denotes whether the method aims for a dialogue task. "Mo" denotes data modality (T – Text, I – Images, B – Box (Position)).

Memory Tool	Level	Taxonomy	Operation	Function	Input/Output	Example Use	Source Type	Access
FAISS (Douze et al., 2024)	Components	Contextual- Unstructured	Consolidation, Indexing and retrieval	Library for fast storage, indexing, and Retrieval of high-dimensional vectors	vector/Index, relevance score	Vector Database-Index a large set of text embeddings and quickly retrieve the most relevant documents for a user's query in a retrieval-augmented generation (RAG) system.	open	[LINK]
Neo4j (Neo4j, 2012)	Components	Contextual- Structured	Consolidation, Indexing, Updating, Retrieval	Native graph database supporting ACID transactions and Cypher query language	Nodes and relationships with properties / Query results via Cypher	Graph Database - Model and retrieve complex relational data for use cases like fraud detection and recommendation engines.	conditional open	[LINK]
BM25 (Robertson et al., 1995)	Components	Contextual- Unstructured	Retrieval	A probabilistic ranking function used in information retrieval to estimate the relevance of documents to a given search query.	Text queries / Ranked list of documents	Enhancing search engine results and document retrieval systems.	open	[LINK]
Contriever (Izacard et al., 2021)	Components	Contextual- Unstructured	Retrieval	An unsupervised dense retriever trained with contrastive learning, capable of retrieving semantically similar documents across languages.	Query text / List of similar documents	High-recall retrieval tasks in multilingual question-answering systems.	open	[LINK]
Embedding Models (e.g. OpenAI embedding (OpenAI, 2025))	Components	Contextual	Consolidation, Retrieval	Techniques that convert text, images, or audio into dense vector representations capturing semantic meaning.	Raw data / Vector embeddings	Text similarity computation, recommendation systems, and clustering tasks.	open	[LINK]

Table 15: Component-Level Tools for Memory Management and Utilization.

Memory Tool	Level	Taxonomy	Operation	Function	Input/Output	Example Use	Source Type	Access
Graphiti (He et al., 2025)	framework	Contextual- Structured	Consolidation, Indexing, Updating, Retrieval	Framework for building and querying temporally-aware knowledge graphs tailored for AI agents in dynamic environments.	Multi-source data / Queryable knowledge graph	Constructing real-time knowledge graphs to enhance AI agent memory.	open	[LINK]
LLamaIndex (Liu, 2022)	framework	Contextual	Consolidation, Indexing, Retrieval	A flexible framework for building knowledge assistants using LLMs connected to enterprise data.	Text / Context- augmented responses	Developing knowledge assistants that process complex data format.	open	[LINK]
LangChain (Chase, 2022)	framework	Contextual	Consolidation, Indexing, Updating, Forgetting, Retrieval	Provides a framework for building context-aware, reasoning applications by connecting LLMs with external data sources.	Input prompts / Multi-step reasoning outputs	Creating complex LLM applications like question-answering systems and chatbots.	open	[LINK]
LangGraph (Inc., 2025)	framework	Contextual- Structured	Consolidation, Indexing, Updating, Forgetting, Retrieval	Constructs controllable agent architectures supporting long-term memory and human-in-the-loop multi-agent systems.	Graph state/ State updates	Building complex task workflows with multiple AI agents.	open	[LINK]
EasyEdit (Wang et al., 2024d)	framework	Parametric	Updating	An easy-to-use knowledge editing framework for LLMs, enabling efficient behavior modification within specific domains.	Edit instructions / Updated model behavior	Modifying LLM knowledge in specific domains, such as updating factual information.	open	[LINK]
CrewAI (Duan and Wang, 2024)	framework	Contextual	Consolidation, Indexing, Retrieval	A platform for building and deploying multi-agent systems, supporting automated workflows using any LLM and cloud platform.	Multi-agent tasks / Collaborative results	Automating workflows across agents like project management and content generation.	open	[LINK]
Letta (Packer et al., 2023)	framework	Contextual- Unstructured	Consolidation, Retrieval	Constructs stateful agents with long-term memory, advanced reasoning, and custom tools within a visual environment.	User interactions / Improved Response	Developing AI agents that learn and improve over time.	open	[LINK]

Table 16: **Framework-Level** Tools for Memory Management and Utilization.

Memory Tool	Level	Taxonomy	Operation	Function	Input/Output	Example Use	Source Type	Access
Mem0 (Taran- jeet Singh, 2024)	Application Layer	Contextual- Unstructured	Consolidation, Indexing, Updating, Retrieval	Provides a smart memory layer for LLMs, enabling direct addition, updating, and searching of memories in models.	User interactions / Personalized responses	Enhancing AI systems with persistent context for customer support and personalized recommendations.	open	[LINK]
Zep (Rasmussen et al., 2025)	Application Layer	Contextual- Structured	Consolidation, Indexing, Updating, Retrieval	Integrates chat messages into a knowledge graph, offering accurate and relevant user information.	Chat logs, business data / Knowledge graph query results	Augmenting AI agents with knowledge through continuous learning from user interactions.	open	[LINK]
Memary (kingjulio823 2025)	Application Layer	Contextual	Consolidation, Indexing, Updating, Retrieval	An open memory layer that emulates human memory to help AI agents manage and utilize information effectively.	Agent tasks / Memory management and utilization	Building AI agents with human-like memory characteristics.	open	[LINK]
Memobase (memodb io, 2025)	Application Layer	Contextual	Consolidation, Indexing, Updating, Retrieval	A user profile-based long-term memory system designed to provide personalized experiences in generative AI applications.	User interactions / Personalized responses	Implementing virtual assistants, educational tools, and personalized AI companions.	open	[LINK]

Table 17: Application Layer-Level Tools for Memory Management and Utilization.

Memory Tool	Level	Taxonomy	Operation	Function	Input/Output	Example Use	Source Type	Access
Me.bot	product	Contextual	Consolidation, Indexing, Updating, Retrieval	AI-powered personal assistant that organizes notes, tasks, and memories, providing emotional support and productivity tools.	User inputs (text, voice) / Organized notes, reminders, summaries	Personal productivity enhancement, emotional support, idea organization.	closed	[LINK]
ima.copilot	Product	Contextual	Consolidation, Indexing, Updating, Retrieval	Intelligent workstation powered by Tencent's Mix Huang model, building a personal knowledge base for learning and work scenarios.	User queries / Customized responses, knowledge retrieval	Enhancing learning efficiency, work productivity, knowledge management.	closed	[LINK]
Coze (Coze, 2024)	Product	Contextual	Consolidation	Enabling multi-agent collaboration across various platforms.	User-defined workflows/ Response	Deployed chatbots, AI agents	closed	[LINK]
Grok (xAI, 2023)	Product	Contextual	Retrieval, Compression	AI assistant developed by xAI, designed to provide truthful, useful, and curious responses, with real-time data access and image generation.	Query / Informative answers, generated images	Answering questions, generating images, providing insights.	closed	[LINK]
ChatGPT (OpenAI, 2022)	Product	Contextual	Consolidation, Retrieval	Conversational AI developed by OpenAI, capable of understanding and generating human-like text based on prompts.	User prompts / Generated text responses	Answering questions, generating images, providing insights.	closed	[LINK]

Table 18: **Product-Level** Tools for Memory Utilization.