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

Anonymous ACL submission

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 memory dynamics. In this survey, we first categorize memory representations into parametric and contextual forms, and then introduce six fundamental memory operations: Consolidation, Updating, Indexing, Forgetting, Retrieval, and Compression. We 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 research.

1 Introduction

Memory is central to LLM-based systems (Wang et al., 2024j), enabling coherent and long-term interaction (Maharana et al., 2024; Li et al., 2024a). While recent work has addressed storage (Zhong et al., 2024), retrieval (Qian et al., 2024; Wang et al., 2025a), and memory-grounded generation (Lu et al., 2023; Yang et al., 2024b; Lee et al., 2024a), cohesive architectural views remain underdeveloped (He et al., 2024c).

Recent surveys have proposed operational views of memory (Zhang et al., 2024f), but most focus narrowly on subtopics such as long-context modeling (Huang et al., 2023b), long-term memory (He et al., 2024c; Jiang et al., 2024b), personalization (Liu et al., 2025), or knowledge editing (Wang et al., 2024g), without offering a unified operational framework. For example, Zhang et al. (2024f) cover only high-level operations such as writing, management, and reading and miss some operations like indexing. More broadly, few surveys define the scope of memory research, analyze technical implementations, or provide practical foundations such as benchmarks and tools.

To address these gaps, we categorize memory into *parametric* and *contextual* types. Parametric memory encodes knowledge implicitly in model parameters (Wang et al., 2024c), while contextual memory stores explicit external information, either structured (Rasmussen et al., 2025) or unstructured (Zhong et al., 2024). Temporally, memory spans both long-term (e.g., multi-turn dialogue, external observations (Li et al., 2024a)) and short-term contexts (Packer et al., 2023). Based on these types, we divide memory operations into *management* and *utilization*. Memory management includes: consolidation (integrating new knowledge into persistent memories (Feng et al., 2024)), indexing (organizing memory for retrieval (Wu et al., 2025a)), updating (modifying memory based on new inputs (Chen et al., 2024b)), and forgetting (removing outdated or incorrect content (Tian et al., 2024)). Memory utilization covers retrieval (accessing relevant memory (Gutiérrez et al., 2024)) and compression (reducing size while preserving key information (Chen et al., 2024b)).

To ground our taxonomy and map key memory-centric research directions, we conduct a pilot study and define four core topics spanning complementary dimensions: (1) **Long-Term Memory** (temporal), covering memory management, utilization, and personalization; (2) **Long-Context Memory** (contextual), focusing on parametric efficiency in extended input handling; (3) **Parametric Memory Modification** (model-internal), including editing, unlearning, and continual learning; and (4) **Multi-Source Memory** (modality/integration), addressing cross-textual (structured/unstructured) integration and multimodal coordination. Based on this taxonomy, we collect and annotate over 30K pa-

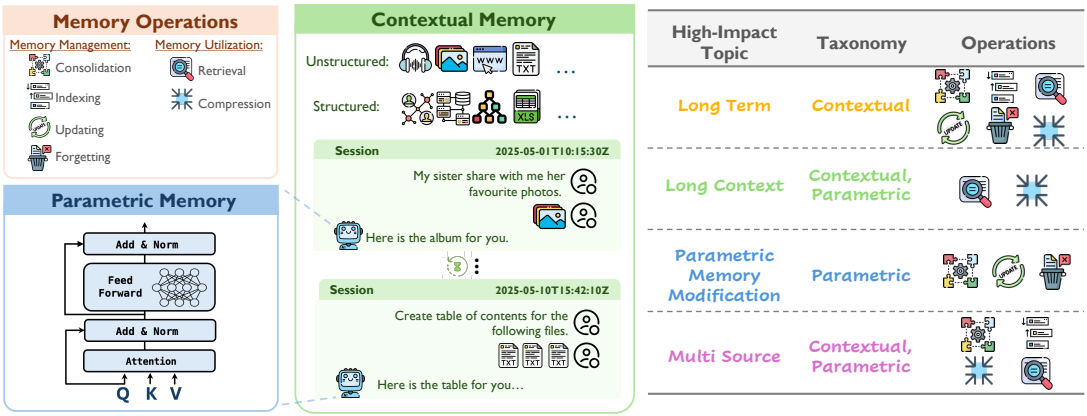

Figure 1: A unified framework of memory Taxonomy, Operations, and High-impact Topics.

pers[1] using a GPT-based relevance scoring pipeline (see Appendix A for details), retaining 3,923 high-relevance papers (score ≥ 8; details in Appendix B). To highlight influential work, we propose the Relative Citation Index (RCI), a time-normalized citation metric inspired by RCR (Hutchins et al., 2016). These papers are systematically analyzed through our unified taxonomy–operations framework (see Table 1 in Appendix).

The remainder of the paper is organized as follows. Section 2 introduces the memory taxonomy and core operations. Section 3 maps high-impact topics to these foundations and summarizes key methods and datasets (Appendix 5). Section 4 outlines practical tools and applications for building memory-enabled AI systems. Section 5 concludes with future directions for memory-centric AI (see Figure 1 for an overview).

2 Memory Foundations

2.1 Taxonomy

From the perspective of memory representation, we divide memory into **Parametric Memory** and **Contextual Memory**, the latter comprising *Unstructured* and *Structured* forms.

Parametric Memory refers to the knowledge implicitly stored within an LLM's internal parameters (Berges et al., 2024; Wang et al., 2024c; Prashanth et al., 2024). Learned during (pre/post-)training, it enables fast, immediate, and context-free access to factual and commonsense knowledge via feedforward computation. This form of long-term memory is persistent and efficient but lacks transparency and is difficult to update selectively in response to new experiences or task-specific contexts.

Contextual Memory denotes explicit, external information that complements an LLM's parameters. *(a) Unstructured Contextual Memory* stores heterogeneous inputs such as text (Zhong et al., 2024), images (Wang et al., 2025a), audio, and video (Wang et al., 2023c), supporting integration across short-term (e.g., current dialogue) and long-term (e.g., user history) contexts (Li et al., 2024a). *(b) Structured Contextual Memory* organizes information into predefined, interpretable formats such as knowledge graphs (Oguz et al., 2022), tables (Lu et al., 2023), or ontologies (Qiang et al., 2023), enabling symbolic reasoning and precise querying. These structures can be transient (built at inference) or persistent (cross-session knowledge bases).

2.2 Operations

Dynamic memory in AI systems relies on operations that govern the information lifecycle and enable effective use during interaction. These fall into two categories: **Memory Management** and **Memory Utilization** (see Figure 1).

2.2.1 Memory Management

Memory management governs how memory is stored, maintained, and pruned over time. It includes four core operations: Consolidation, Indexing, Updating, and Forgetting, all reflecting the temporal dynamics of memory.

Consolidation (Squire et al., 2015) refers to transforming m short-term experiences $\mathcal{E}_{[t,t+\Delta_t]} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_m)$ elapsing between t and $t + \Delta_t$ into persistent memory $\mathcal{M}_{t+\Delta_t}$. This involves encoding interaction histories (e.g., dialogues) into durable forms such as parameters (Wang et al.,

[1] From NeurIPS, ICLR, ICML, ACL, EMNLP, and NAACL (2022–2025).

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).

$$\mathcal{M}_{t+\Delta_t} = \texttt{Consolidate}(\mathcal{M}_t, \mathcal{E}_{[t,t+\Delta_t]}) \quad (1)$$

Indexing (Maekawa et al., 2023) builds auxiliary codes ϕ (e.g., entities, embeddings (Wu et al., 2025a)) to support efficient and structured memory retrieval, enabling temporal (Maharana et al., 2024) and relational traversal (Mehta et al., 2022) across memories. It supports scalable retrieval across symbolic, neural, and hybrid memory systems.

$$\mathcal{I}_t = \texttt{Index}(\mathcal{M}_t, \phi) \quad (2)$$

Updating (Kiley and Parks, 2022) reactivates existing memory representations in $\mathcal{M}_t$ and modifies them with new knowledge $\mathcal{K}_{t+\Delta_t}$. Updating parametric memory 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.

$$\mathcal{M}_{t+\Delta_t} = \texttt{Update}(\mathcal{M}_t, \mathcal{K}_{t+\Delta_t}) \quad (3)$$

Forgetting (Davis and Zhong, 2017; Wang et al., 2009) refers to selectively removing memory content $\mathcal{F}$ from $\mathcal{M}_t$ that is outdated or harmful. In parametric memory, this is achieved via unlearning techniques (Jia et al., 2024a; Li et al., 2025). In contextual memory, forgetting involves time-based deletion (Zhong et al., 2024) or semantic filtering (Wang et al., 2024e).

$$\mathcal{M}_{t+\Delta_t} = \texttt{Forget}(\mathcal{M}_t, \mathcal{F}) \quad (4)$$

Despite its benefits, forgetting poses security risks via persistent malicious edits. (see Section 5).

2.2.2 Memory Utilization

Memory utilization refers to how memory is accessed and used during inference, comprising two operations: Retrieval and Compression.

Retrieval selects relevant memory fragments $m_{\mathcal{Q}}$ in response to inputs $\mathcal{Q}$ (ranging from textual queries (Du et al., 2024), multi-modal queries or multi-turn dialogues (Wang et al., 2025a; Zhou et al., 2024)). Memory fragments are scored with a function *sim()* with those above a threshold τ deemed relevant. Retrieval targets include memory from multiple sources (Tan et al., 2024b), modalities (Wang et al., 2025a), or even parametric representations (Luo et al., 2024) within LLMs.

$$\texttt{Retrieve}(\mathcal{M}_t, \mathcal{Q}) = m_{\mathcal{Q}} \in \mathcal{M}_t \quad \text{with } \text{sim}(\mathcal{Q}, m_{\mathcal{Q}}) \geq \tau \quad (5)$$

Compression improves efficiency by reducing memory size with compression ratio α, either before input (e.g., filtering long contexts (Yu et al., 2023)) or after retrieval (e.g., summarizing retrieved content (Xu et al., 2024a; Safaya and Yuret, 2024)). Unlike memory consolidation, which summarizes information during memory construction (Zhong et al., 2024), compression focuses on reducing memory for inference (Lee et al., 2024a).

$$\mathcal{M}_t^{comp} = \texttt{Compress}(\mathcal{M}_t, \alpha) \quad (6)$$

3 From Operations to Primary Topics

This section analyzes how real-world systems manage and utilize memory through core operations. We examine four key research topics introduced in Section 1, guided by the framework in Figure 1, using the Relative Citation Index (RCI)—a time-adjusted metric normalizes citation counts by publication age (Appendix B)—to highlight influential work. RCI surfaces emerging trends and enduring contributions across memory research. Figure 6 shows the architectural landscape of these topics.

3.1 Long-term Memory

Long-term memory refers to the persistent storage of information acquired through interactions such as multi-turn dialogues. It enables memory **management**, **utilization**, and **personalization** across extended interactions. This section focuses on contextual long-term memory. See in Appendix Tables 3 for representative datasets and Tables 7 and 8 for representative approaches.

Memory Management. A core component of long-term memory systems, memory management includes consolidation, indexing, updating, and forgetting. ***Consolidation*** turns short-term inputs into persistent memory via summarization, salient extraction, or temporal modeling (Lu et al., 2023; Zhong et al., 2024; Hou et al., 2024; Wang et al., 2025c; Park et al., 2025). ***Indexing*** ensures efficient access, using graph-based, timestamped, or

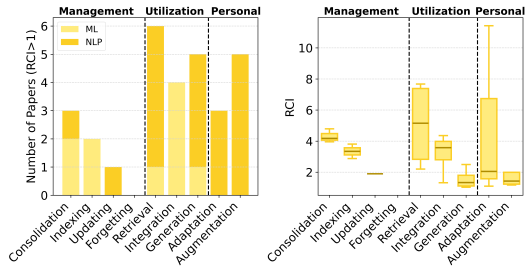

Figure 2: Publication statistic of highlighted papers (RCI > 1) discussed in long-term memory.

timeline-aware structures (Gutiérrez et al., 2024; Wu et al., 2025a; iunn Ong et al., 2025). ***Updating*** modifies memory content through recursive summarization, selective deletion, or reflective rewriting, sometimes guided by external feedback (Dalvi Mishra et al., 2022; Bae et al., 2022; Sun et al., 2024; Wang et al., 2025b). ***Forgetting***, though less explored, plays a critical role in safety and compliance, implemented via passive decay in external memory (Zhong et al., 2024; Chen et al., 2024b). Parametric unlearning is discussed separately in Section 3.3.

Memory Utilization. A core process in long-term memory systems is memory utilization, involves retrieving, integrating, and applying memory during inference. ***Retrieval*** can be query-centered, memory-centered, or event-centered, with advanced strategies like multi-hop traversal and graph-based evolution (Xu et al., 2021; Jiang et al., 2023b; Jang et al., 2024; Wu et al., 2025a; Du et al., 2024; Maharana et al., 2024; Gutiérrez et al., 2024; Qian et al., 2024). ***Integration*** is either static—directly merging retrieved memory with context—or dynamic, where memory evolves through interaction (Chen et al., 2024a; Li et al., 2024h; Hou et al., 2024; Zheng et al., 2024). Retrieved memory further guides grounded ***generation*** via reflection, feedback, and long-context alignment (Tandon et al., 2021; Lu et al., 2023; Li and Qiu, 2023; Li et al., 2024i; Chen et al., 2024b; Lee et al., 2024b).

Personalization. Essential for user-adaptive behavior, personalization combines model adaptation and memory augmentation. ***Adaptation*** encodes user preferences via fine-tuning or lightweight modules like prefix encoders, adapters, or latent embeddings (Liu et al., 2023c; Tang et al., 2023a; Tan et al., 2024d). Dual-memory systems such as MaLP model both long- and short-term traits (Zhang et al., 2023b). ***Augmentation*** retrieves structured profiles, unstructured histories, or hybrid memory from persistent agents (Dutt et al., 2022; Fu et al., 2022; Salemi et al., 2023; Huang et al., 2024a; Zhong et al., 2024; Li et al., 2024a). Despite scalability, most approaches remain passive, revealing challenges in building adaptive and proactive personalization.

Discussion. *1) Static Memory Limit Evaluation.* Most current evaluations focus on retrieval and generation accuracy in factual Question Answering or multi-turn dialogue (Yang et al., 2024c; Salama et al., 2025; Wu et al., 2025a; Maharana et al., 2024), often assuming static memory and overlooking operations like updating, selective retention, and cross-session continuity. This static view limits our understanding of how models manage memory over time. *2) Gap Between Retrieval and Generation.* While benchmarks such as LoCoMo (Maharana et al., 2024) and MemoryBank (Zhong et al., 2024) incorporate longer contexts, they fail to account for temporal drift, source inconsistency, and memory reliability, leading to a disconnection between retrieval scores and generation quality under noisy or distant conditions (see Figure 13). *3) Personalization and Planning Require Evolving Memory.* Recent work has explored personalization through profile retrieval and agent-based modeling of long-term user behavior (Salemi et al., 2023; Dutt et al., 2022; Fu et al., 2022; Li et al., 2024a), but often assumes static profiles and offers limited evaluation of how memory consistency, user adaptation, and planning based on evolving memory unfold across sessions.

As shown in Figure 2, retrieval and generation dominate recent literature, especially in NLP. Core operations like consolidation and indexing receive more focus in ML, while forgetting remains underexplored. Personalization is largely limited to NLP due to practical relevance. In terms of citation impact, consolidation, retrieval, and integration play key roles—driven by advances in memory-aware fine-tuning, summarization, retrieval-augmented generation, and prompt fusion.

3.2 Long-context

Managing vast quantities of multi-sourced external memory in conversational search presents significant challenges in long-context language understanding. These challenges can be broadly categorized into **Parametric Efficiency** and **Contextual Utilization**. In this section, we review efforts

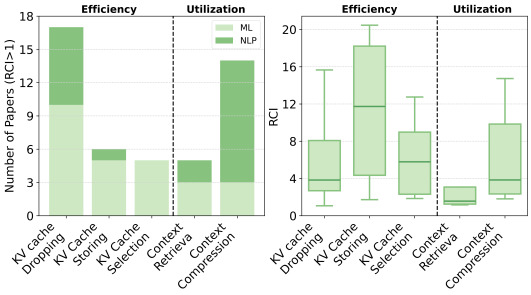

Figure 3: Publication statistic of highlighted papers (RCI > 1) discussed in long-context memory.

made in handling these challenges. Representative datasets and methods are reviewed in Tables 4, 9 and 10 in Appendix.

Parametric Efficiency. Key-Value (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, with static approaches (Xiao et al., 2024; Han et al., 2024) dropping KV cache with fixed pattern, dynamic approaches dropping KV cache concerning the query (Zhang et al., 2023c; Ge et al., 2024; Chen et al., 2024c), or the model behavior (Liu et al., 2023d; Li et al., 2024g; Yang et al., 2024a; Yao et al., 2024a). ***KV Cache Storing Optimization*** considers the potential information loss when removing KV cache by preserving the entire KV cache at a smaller footprint, achieved through compressing less important cache entries into low-rank representations (Dong et al., 2024), or dynamically quantize KV cache to reduce memory allocation (Liu et al., 2024f; Zhao et al., 2024c; Hooper et al., 2024; Sheng et al., 2023)). ***KV Cache Selection*** refers to selectively loading required KV cache to speed up the inference, which focuses on memory retrieval (Wu et al., 2022a; Tworkowski et al., 2023; Tang et al., 2024).

Contextual Utilization. Apart from optimizing language models to obtain long-context abilities, optimizing contextual memory utilization raises another important challenge. ***Context Retrieval*** aims to enhance LLM's ability in identifying and locating key information from the contextual memory. Graph-based approaches (Li et al., 2024d) decompose documents into graph structures for effective context selection. Token-level methods (Yu et al., 2023; Zhang et al., 2024c) selecting tokens deemed most important while fragment-level methods (Zhu et al., 2025) perform context selection at the fragment level. Training-based approaches (He et al., 2024b; An et al., 2024c) train LLMs with specialized data to improve the context selection ability. ***Context Compression*** utilizes memory compression operation to optimize contextual memory utilization. Soft prompt compression (Chevalier et al., 2023; Cheng et al., 2024) focuses on compressing chunks of input tokens into continuous vectors. Hard prompt compression directly compress long input chunks into shorter natural language chunks by dropping uninformative tokens (Li et al., 2023) or chunks (Fei et al., 2024), abstracting the key information to summarize the context (Jiang et al., 2023a, 2024a; Pan et al., 2024), or combining dropping and abstracting (Liu et al., 2023a).

Discussion: *1) Compression vs. Performance Trade-off.* Yuan et al. (2024) propose an universal benchmarking on different compression strategies (Figure 14), showcasing that KV cache storage optimization methods achieve best trade-off between effectiveness and efficiency. In contrast, KV cache dropping methods are more flexible but less effective. In the other hand, compressing the contextual memory are less effective compared with compressing the parametric memory. *2) Lost in Context.* Despite efforts to extend context length to millions of tokens (Ding et al., 2023), long-context LLMs have been found to miss crucial information in the middle of the context (Liu et al., 2024d; Ravaut et al., 2024). In addition, though higher recall can be obtained with larger retrieval set, irrelevant information will mislead LLMs and harm the generation quality (Shi et al., 2023; Jin et al., 2025).

In publication trend perspective, Figure 3 shows that the NLP community focus more on the utilization aspect with contextual memory, while the ML community dedicate more effort on efficiency processing with parametric memory. From an RCI perspective, KV cache storage optimization dominates discussions on this topics. This dominance stems from their optimal balance efficiency and effectiveness, as well as their general compatibility with other long-context methods.

3.3 Parametric Memory Modification

Modifying parametric memory, which is encoded knowledge within the LLM parameters, is crucial for dynamically adapting stored memory. Existing methods for parametric memory modification

can be grouped into three categories: **Editing**, **Unlearning**, and **Continual Learning**. Representative datasets and methods are reported in Tables 5, 11, 12, and 13.

Editing Editing refers to updating specific knowledge in a model's parametric memory without full retraining. One prominent line of work involves directly modifying model weights. A dominant strategy is ***Locating-then-Editing*** (Meng et al., 2022a, 2023; Mela et al., 2024; Huang et al., 2024b; Fang et al., 2025), which first identifies and then updates the relevant parameters. Another approach is ***meta-learning*** (De Cao et al., 2021; Mitchell et al., 2022a; Tan et al., 2024a; Li et al., 2024e; Zhang et al., 2024d), where an auxiliary network learns to generate efficient weight updates. Some methods avoid altering the original weights altogether: ***Prompt***-based methods steer the model through in-context prompts (Zheng et al., 2023; Zhong et al., 2023), while ***Additional Parameter*** methods store updates externally in learnable modules (Mitchell et al., 2022c; Dong et al., 2022; Wang et al., 2024c,i; Das et al., 2024).

Unlearning Unlearning aims to remove specific knowledge from a model while preserving unrelated information. ***Additional Parameter*** methods introduce modules such as logit difference units (Ji et al., 2024) or dedicated unlearning layers (Chen and Yang, 2023). ***Prompt***-based approaches either modify the input directly (Liu et al., 2024b) or apply in-context learning techniques (Pawelczyk et al., 2024). ***Locating-then-Unlearning*** methods (Jia et al., 2024a; Tian et al., 2024; Wu et al., 2023) identify and suppress the memory responsible for undesired behavior. Finally, ***Training Objective***-based methods (Wang et al., 2025d; Liu et al., 2024e; Jia et al., 2024b; Yao et al., 2024b) revise the loss function to encourage forgetting.

Lifelong learning Lifelong learning (Wang et al., 2024b) enables long-term memory retention by mitigating catastrophic forgetting. ***Regularization-based learning*** (Feng et al., 2024; Wang et al.; Kirkpatrick et al., 2017; Wu et al., 2024) constrains updates to important weights to preserve prior knowledge. ***Replay-based learning*** (Mehta et al., 2022) reinforces memory by reintroducing past samples, supporting the integration of retrieved or historical knowledge. ***Interactive learning***, as in LifeSpan Cognitive System (Wang et al., 2024j), allows agents to acquire and consolidate memory through real-time experience, offering insights into continual parametric encoding.

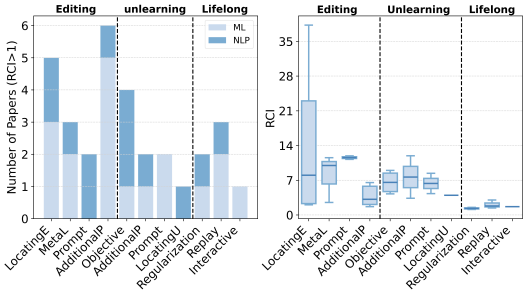

Figure 4: Publication statistic of highlighted papers (RCI > 1) discussed in parametric memory modification. LocatingU, LocatingE and AdditionalP refer to locating-then-editing, locating-then-unlearning and additional parameters, respectively.

Discussion: *1) Editing Still Requires Precise Control.* As shown in Figure 16, current editing methods perform poorly on the ZsRE benchmark due to low specificity, underscoring the challenge of making precise edits without affecting unrelated information. *2) Unlearning needs a More Challenging and Realistic Benchmark.* As detailed in Figure 15, current unlearning methods already achieve high scores on TOFU, suggesting that it may not be challenging enough. This indicates a need for new unlearning benchmarks that go beyond the unlearning of specific entities. *3) Scalability Remains Underexplored.* Most methods (Figure 17) attempted no more than 5,000 edits, with limited exploration of sequential unlearning. Non-prompt approaches (Figure 18) are costly and limited to small models (< 20B). The link between model size and its capacity for edits or unlearning remains unknown. Enabling efficient, scalable editing and unlearning is a key open challenge.

In publication trend perspective, Figure 4 shows that research focuses mainly on editing, followed by unlearning, with less attention to lifelong learning. Editing has higher impact, while unlearning methods—especially those using additional parameters—are gaining interest. This suggests a shift toward post-deployment model adjustment, with lifelong learning still underexplored.

3.4 Multi-source Memory

Multi-source memory is crucial for real-world AI, encompassing both parametric memory and contextual memory. These memories support reasoning across short-term context and long-term user history or domain knowledge. Key challenges include **cross-textual integration** and **multi-modal coor-**

dination across these heterogeneous sources.

Cross-textual Integration. Text-based memory integration requires factual consistency and cross-domain ***Reasoning***. Recent efforts combine structured and unstructured sources (Hu et al., 2023; Wang et al., 2024f; Xu et al., 2024c) or merge parameterized and retrieved content (Nogueira dos Santos et al., 2024; Wang et al., 2025e). However, ***Conflicts*** often emerge when merging heterogeneous inputs. Techniques like RKC-LLM (Wang et al., 2023b) and BGC-KC (Tan et al., 2024b) detect inconsistencies and propose source-aware trust mechanisms, yet remain limited in dynamic or multi-session settings.

Multi-Modal Coordination. In multi-modal scenarios, fusion and retrieval are central to memory usage. Unified embedding spaces (e.g., UniTransSeR (Ma et al., 2022), PaLM-E (Driess et al., 2023)) enable short-term cross-modal ***Fusion***, while approaches like LifelongMemory (Wang et al., 2023c) and MA-LMM (He et al., 2024a) accumulate long-term cross-modal knowledge. ***Retrieval*** remains embedding-based (e.g., CLIP (Radford et al., 2021), QwenVL (Bai et al., 2023)), IGSR (Wang et al., 2025a) with limited capacity for reasoning or leveraging underexplored signals like audio. Future systems must bridge this retrieval–reasoning gap and support persistent, multi-modal memory grounded in temporal dynamics.

Discussion. *1) Conflict-aware Reasoning Needed.* Cross-textual memory integration is shifting from symbolic querying to generative reasoning. Early work relied on structured symbolic memory (Wu et al., 2022b; Hu et al., 2023), while later work introduced unstructured retrieval and attention-based inference (Li et al., 2024i; Wang et al., 2025e), still treating memory as static. Recent systems embed memory into reasoning (Xu et al., 2024c; Michelman et al., 2025), but often merge retrieved and parametric content without resolving semantic conflicts, leading to hallucinations (Zhou et al., 2023; Tan et al., 2024c). Some efforts apply epistemic calibration or multi-step resolution (Wang et al., 2023b; Xu et al., 2024b), but remain limited in scope. *2) Temporal and Structured Integration Are Converging.* Time-aware fusion and retrieval have become common in recent multi-modal memory models for long-horizon reasoning (Figure 21), highlighting a shift toward temporal and operational integration (Wang et al., 2023c; He et al., 2024a; Zhou et al., 2024). General-purpose solutions such as joint embedding and prompt-level fusion are commonly adopted, while more task-specific approaches leverage identifier-based retrieval and graph-based coordination to achieve precise integration (Nguyen et al., 2023; Li et al., 2024f). *3) Operational Scalability Remains Underexplored.* Beyond retrieval, operations like indexing, updating, and compression are increasingly central. Emerging system adopt self-maintaining memory across sessions (Glocker et al., 2025; Xiao et al., 2025), moving from passive storage to actively managed memory, especially in multi-source contexts.

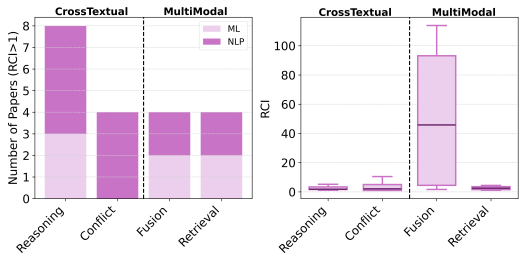

Figure 5: Publication statistic of highlighted papers (RCI > 1) discussed in multi-source memory.

As shown in Figure 5, cross-textual reasoning dominates by publication volume, reflecting its central role in multi-source integration. Fusion research, particularly work driven by CLIP (Radford et al., 2021), achieves the highest RCI and demonstrates strong influence on multi-modal learning. In contrast, progress in multi-modal retrieval remains limited, and conflict resolution is still narrowly explored within NLP. The overall slowdown suggests a transition toward consolidation in these areas.

4 Memory *In Practice*

Applications. Memory-centric systems enable knowledge retrieval, personalization, and long-horizon planning in real-world scenarios. **Knowledge-centric systems** encode general knowledge in model weights (Chen et al., 2021a; Yang et al., 2023; Bi et al., 2023), supporting medical, legal, or financial assistants requiring static expertise. **User-centric systems** leverage contextual memory to model preferences and history (Li et al., 2024a; Qin et al., 2025; Hong et al., 2023), powering mental health chatbots and personalized tutoring. **Task-oriented agents** use structured memory for session continuity and long-range reasoning (Xu et al., 2025), such as project assistants tracking meeting notes.

Multi-modal systems (OpenAI, 2023) integrate all memory types to enable coherent interaction in settings like in-car copilots or medical tools.

Products. AI companions (e.g., Replika (Luka, Inc., 2025)), recommender systems (e.g., Amazon (Linden et al., 2003)), and virtual assistants (e.g., Me.bot, Tencent ima.copilot (Coze, 2024; xAI, 2023)) exemplify user-centric memory design. Task-oriented tools such as ChatGPT, Grok, GitHub Copilot, Coze, and CodeBuddy (OpenAI, 2022; xAI, 2023; GitHub and OpenAI, 2021; Coze, 2024; Zhao et al., 2024a) showcase structured contextual memory in real-world deployment.

Tools. A layered memory ecosystem has emerged to support these applications. Core *components* include vector stores (e.g., FAISS (Douze et al., 2024)), graph databases (e.g., Neo4j (Neo4j, 2012)), and LLMs (e.g., LLaMA (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), DeepSeek (Liu et al., 2024a)). Retrieval tools such as BM25 (Robertson et al., 1995), Contriever (Izacard et al., 2021), and OpenAI embeddings (OpenAI, 2025) enable semantic access. On top of these, *frameworks* like LangChain (Chase, 2022), LlamaIndex (Liu, 2022), and Graphiti (He et al., 2025) provide modular pipelines. Mid-layer *orchestration systems* such as Zep (Rasmussen et al., 2025), Mem0 (Taranjeet Singh, 2024), and Memary (kingjulio8238, 2025) manage memory lifecycle and temporal consistency. Tool details are listed in Tables 16–19 in Appendix.

5 Challenge and Future Direction

Designing memory-centric AI requires addressing core limitations and emerging demands. Guided by RCI analysis and trends, we outline key challenges shaping future memory research.

Unified evaluation is needed to address consistency, personalization, and temporal reasoning in long-term memory. Existing benchmarks rarely assess core operations such as consolidation, updating, retrieval, and forgetting in dynamic, multi-session settings. This gap contributes to the retrieval–generation mismatch, where retrieved content is often outdated, irrelevant, or misaligned due to poor memory maintenance. Addressing these issues requires temporal reasoning, structure-aware generation, and retrieval robustness along with systems supporting personalized reuse and adaptive memory management across sessions.

Long-context Processing: Efficiency vs. Expressivity. Scaling memory length exacerbates trade-offs between computational cost and modeling fidelity. Techniques like KV cache compression and recurrent memory reuse offer efficiency, but risk information loss or instability. At the same time, reasoning over complex environments, especially in multi-source or multi-modal settings, requires selective context integration, source differentiation, and attention modulation. Bridging these demands mechanisms that balance contextual bandwidth with task-specific relevance and stability.

While promising, parametric memory modification requires further research to improve control, erasure, and scalability. Current editing methods often lack specificity, while unlearning benchmarks like TOFU may be too simple to reveal real limitations. Most approaches do not scale beyond a few thousand edits or support models over 20B parameters. Additionally, lifelong learning is still underexplored despite its potential. Future work should develop more realistic benchmarks, improve efficiency, and unify editing, unlearning, and continual learning into a cohesive framework.

Multi-source Integration: Consistency, Compression, and Coordination. Modern agents rely on heterogeneous memory—structured knowledge, unstructured histories, and multi-modal signals—but face redundancy, inconsistency, and source ambiguity. These arise from misaligned temporal scopes, conflicting semantics, and missing attribution, particularly across modalities. Addressing them requires conflict resolution, temporal grounding, and provenance tracking. Efficient indexing and compression are also essential for scalability and interpretability in multi-session settings.

Beyond these core areas, several cross-cutting frontiers are emerging: **spatio-temporal memory**, which captures evolving relational dynamics over time; **unified memory representation**, bridging parametric and contextual spaces; **lifelong learning**, balancing plasticity and stability across memory types; **multi-agent memory**, enabling decentralized synchronization and coordination; **biological inspirations**, including dual-memory systems and hierarchical abstraction; and **memory safety**, ensuring robust unlearning and secure retention under adversarial conditions.

These challenges require systems capable not only of retaining information but also of doing so responsibly, efficiently, and adaptively.

Limitation

Our paper selection primarily focus on memory-centric research, and articles from related but tangential fields are not systematically included or analyzed. Additionally, limiting the scope to the six top conferences may restrict the range of accessible papers reviewed in this study. To mitigate this limitation, we have included additional reviews of highly relevant papers beyond these conferences, including preprints. Apart from this, given the breadth of the reviewed topics and the extensive number of memory-based works, some influential studies may still be missing. To minimize such omissions as much as possible, we utilize an RCI-based approach to ensure that most of the highly influential works are included and discussed in this paper.

Ethics Statement

This study is a literature-based survey and does not involve human subjects, personal data, or experiments requiring ethical approval. All referenced works are publicly available and properly cited. While we mention several commercial systems and products as part of our analysis, we have no affiliations with or financial interests in any of the companies or organizations discussed.

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*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024a. 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.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024b. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024c. Make your llm fully utilize the context. In *Advances in Neural Information Processing Systems*, volume 37, pages 62160–62188. Curran Associates, Inc.

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.

Vincent-Pierre Berges, Barlas Oğuz, Daniel Haziza, Wen tau Yih, Luke Zettlemoyer, and Gargi Ghosh. 2024. Memory layers at scale.

Sheng Bi, Zhiyao Zhou, Lu Pan, and Guilin Qi. 2023. Judicial knowledge-enhanced magnitude-aware reasoning for numerical legal judgment prediction. *Artificial Intelligence and Law*, 31(4):773–806.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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. 2021b. 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*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846, Singapore. Association for Computational Linguistics.

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.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens.

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*.

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*.

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.

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 Representations*.

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*.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2024. Extending context window of large language models via semantic compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5169–5181, Bangkok, Thailand. Association for Computational Linguistics.

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.

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*.

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*.

GitHub and OpenAI. 2021. Github copilot: Your ai pair programmer. `https://github.com/features/copilot`. Accessed May 2025.

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

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*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

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.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: 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 position-agnostic 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*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

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. 2024a. Learning retrieval augmentation for personalized dialogue generation. *arXiv preprint arXiv:2406.18847*.

Xiusheng Huang, Yequan Wang, Jun Zhao, and Kang Liu. 2024b. Commonsense knowledge editing based on free-text in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14870–14880.

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*.

B. Ian Hutchins, Xin Yuan, James M. Anderson, and George M. Santangelo. 2016. Relative citation ratio (rcr): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology*, 14(9):1–25.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,

pages 4276–4292, Miami, Florida, USA. Association for Computational Linguistics.

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.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

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. 2024. Dialsim: A real-time simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*.

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.

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.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024a. 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. 2024b. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.

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*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024b. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

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. 2024c. 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. 2024d. 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-of-thought enables chatgpt to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, Singapore. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024e. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18564–18572.

Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024f. 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.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024g. 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. 2024h. 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. 2024i. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.

Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

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. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023a. TCRA-LLM: Token compression retrieval augmented large language model for inference cost reduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9796–9810, Singapore. Association for Computational Linguistics.

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. 2023b. 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. 2023c. 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*.

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.

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*.

Luka, Inc. 2025. Replika: The ai companion who cares. https://replika.com/. Accessed: 2025-05-14.

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)*.

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. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing 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 Representations*.

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

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. Memory-based 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.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore. 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. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

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.

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.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. Meminsight: Autonomous memory augmentation for llm agents. *arXiv preprint arXiv:2503.21760*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

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.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

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*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. 2015. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766.

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.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024a. Massive editing for large language models via meta learning. In *The Twelfth International Conference on Learning Representations*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024b. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv e-prints*, pages arXiv–2401.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024c. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024d. 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: Query-aware 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. ♫ MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Szymon Tworkowski, Konrad Staniszewski, Mikoł aj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Mił oś. 2023. Focused transformer: Contrastive training for context scaling. In *Advances in Neural Information Processing Systems*, volume 36, pages 42661–42688. Curran Associates, Inc.

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.

Kewen Wang, Zhe Wang, Rodney Topor, Jeff Z. Pan, and Grigoris Antoniou. 2009. Concept and role forgetting in alc ontologies. In *Proceedings of the 8th International Semantic Web Conference (ISWC2009)*, volume 5318.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023a. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*.

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. 2023b. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Ying Wang, Yanlai Yang, and Mengye Ren. 2023c. 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)*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024k. 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. 2025a. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei Wu, Zhuoshi Pan, Chao Wang, Liyi Chen, Yunchu Bai, Tianfu Wang, Kun Fu, Zheng Wang, and Hui Xiong. 2025b. 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. 2024. Mitigating catastrophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization. In *Forty-first International Conference on Machine Learning*.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022a. Memorizing transformers. *arXiv preprint arXiv:2203.08913*.

Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2022b. 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*.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. 2025. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RECOMP: 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. MultiInstruct: 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.

Dongjie Yang, Xiaodong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. 2024a. PyramidInfer: Pyramid KV cache compression for high-throughput LLM inference. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3258–3270, Bangkok, Thailand. 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. 2024b. $memory^3$: Language modeling with explicit memory. *arXiv preprint arXiv:2407.01178*.

Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, et al. 2024c. Memory3: Language modeling with explicit memory. *arXiv preprint arXiv:2407.01178*.

John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, Diyi Yang, Sida Wang, and Ofir Press. 2025. SWE-bench multimodal: Do AI systems generalize to visual software domains? In *The Thirteenth International Conference on Learning Representations*.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

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*.

Yao Yao, Zuchao Li, and Hai Zhao. 2024a. SirLLM: Streaming infinite retentive LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2611–2624, Bangkok, Thailand. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. 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.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. 2024. KV cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4623–4648, Miami, Florida, USA. 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*.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2023b. 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*.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024c. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations*.

Taolin Zhang, Qizhou Chen, Dongyang Li, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, et al. 2024d. 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.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024e. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024f. 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. 2023c. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wayne Xin Zhao, Yusheng Wang, Yujia Yuan, Qitian Xiao, Yichong He, Jingyuan Zhang, and Ji-Rong Wen. 2024a. Codebuddy: Teaching large language models to write better code via self-improvement feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv preprint arXiv:2403.09161.

Wenting Zhao, Ye Liu, Tong Niu, Yao Wan, Philip 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, Mexico City, Mexico. 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: Low-bit 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.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

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*.

A GPT-based Pipeline Selection

To facilitate large-scale relevance filtering aligned with our taxonomy, we design a GPT-based scoring pipeline to evaluate the alignment between paper abstracts and predefined task definitions (Table 2). Each abstract is paired with a corresponding task definition and scored on a 1–10 scale by the model, with a threshold of ≥ 8 used to retain high-relevance papers for further analysis. We adopt **GPT-4o-mini** as the scoring backbone due to its favorable trade-off between performance and efficiency. Despite its relatively lightweight architecture, GPT-4o-mini demonstrates strong zero-shot reasoning capabilities, making it a cost-effective and sufficiently accurate choice for abstract-level topic relevance estimation across a corpus of over 30,000 papers. The exact prompt format used in this evaluation process is illustrated in Figure 10.

B Relative Citation Index

In this work, we identify impactful works by Relative Citation Index (RCI) metric inspired by the RCR metrics (Hutchins et al., 2016), which estimate the expected citations with respect to publication age to prevent bias between original citations from different publication dates. The age A_i of a paper p_i is computed as:

$$A = T - Year_i \tag{7}$$

, 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 in three different way, which are:

linear model:

$$C_i = \beta + \alpha A_i \tag{8}$$

exponential model:

$$C_i = \exp(\beta + \alpha A_i) \tag{9}$$

log-log regression model:

$$\log(C_i + 1) = \beta + \alpha \log A_i + \epsilon_i \tag{10}$$

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 topics discussed in the paper, using the prompt shown in Figure 10. We pick all the papers with score equal and higher than 8 and collect their publication date and citation numbers from Semantic Scholar API[2]. 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 $\hat{\beta}$ and $\hat{\alpha}$ accordingly[3]. Figure 7 shows the estimated age-citation model, where we can find that the log-log regression model best fit the data, which almost perfectly fitting the median citation with respect to publication age. In addition, log-log regression model grantees that the expected citation equals 0 when a paper is freshly released, which follows the intuition. Thus, we pick log-log regression model to compute the expected citation for next step[4], and we are able to obtain the expected citation number $\hat{C}_i$ of paper p_i with age A_i as:

$$\hat{C}_i = \exp(\hat{\beta}) A_i^{\hat{\alpha}} \tag{11}$$

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

$$RCI_i = \frac{C_i}{\hat{C}_i} \tag{12}$$

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

[2] https://www.semanticscholar.org/product/api

[3] Noted 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.

[4] The estimation is: $\hat{\beta} = 1.878$, $\hat{\alpha} = 1.297$

Operations	Parametric	Contextual	
		Structured	Unstructured
Consolidation	Continual Learning, Personalization	Management, Personalization	Management, Personalization
Indexing	Utilization	Utilization, Management, Personalization	Utilization, Management, Personalization, Multi-modal Coordination
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	Utilization, Personalization, Contextual Utilization, Multi-modal Coordination
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.

C RCI-Driven Analysis of Topic Impact

In this study, we leverage both RCI and publication volume trends to gain a clearer understanding of the development and influence of various memory-related research topics. As shown in Figure 8, boxplots illustrate the distribution of median Relative Citation Index (RCI) values across topics by year. Notably, 2023 stands out as a pivotal year following the emergence of large language models (LLMs), with a surge in both the quantity and quality of publications related to long-context and parametric memory, suggesting that these areas were directly shaped by the advancement of LLMs. In contrast, long-term memory and multi-source memory maintained relatively stable average impact levels, indicating continued activity without the emergence of disruptive or field-defining work during that period.

Figure 9 visualizes the temporal trends in publication volume and median RCI for each topic. All topics experienced notable growth in publication counts, with long-context in particular expanding from one of the least represented topics before 2022 to the most prominent by 2024—largely driven by the rise of LLMs. Furthermore, the RCI of long-term memory has shown a steady increase, reflecting a growing body of valuable work in that domain. By contrast, other topics witnessed a noticeable decline in RCI medians after 2023, though their influence levels remained comparable to those seen prior to 2022. These patterns collectively underscore the substantial impact of large models in catalyzing progress across memory-related research, especially in the areas of long-context and parametric memory.

D Chord Analysis of Interactions Among Memory Types, Operations, Topics, and Venues

We present a chord-based analysis of memory research from two perspectives: (1) the interactions among memory types, operations, and topics, and (2) their distribution across major ML and NLP conference venues.

D.1 Memory Interactions Across Types, Operations, and Topics

To intuitively analyze the strength of connections between memory types, operations, and research topics, we examine 132 method-focused papers with an RCI ≥ 1 and generate a final chord diagram (as shown in Figure 11) based on the analysis.

From the perspective of memory types, research predominantly focuses on parametric memory and contextual unstructured memory, with most work centered on compression, retrieval, forgetting, and updating. In contrast, contextual structured memory is relatively underexplored, likely because LLMs are optimized for sequential text and perform less effectively on structured inputs.

From the operation perspective, compression and retrieval are the most frequently studied, while indexing receives comparatively less attention. This is largely because most existing works focus on the use of memory, where retrieval and compression are two fundamental operations. In the case of consolidation, most studies refer to storing

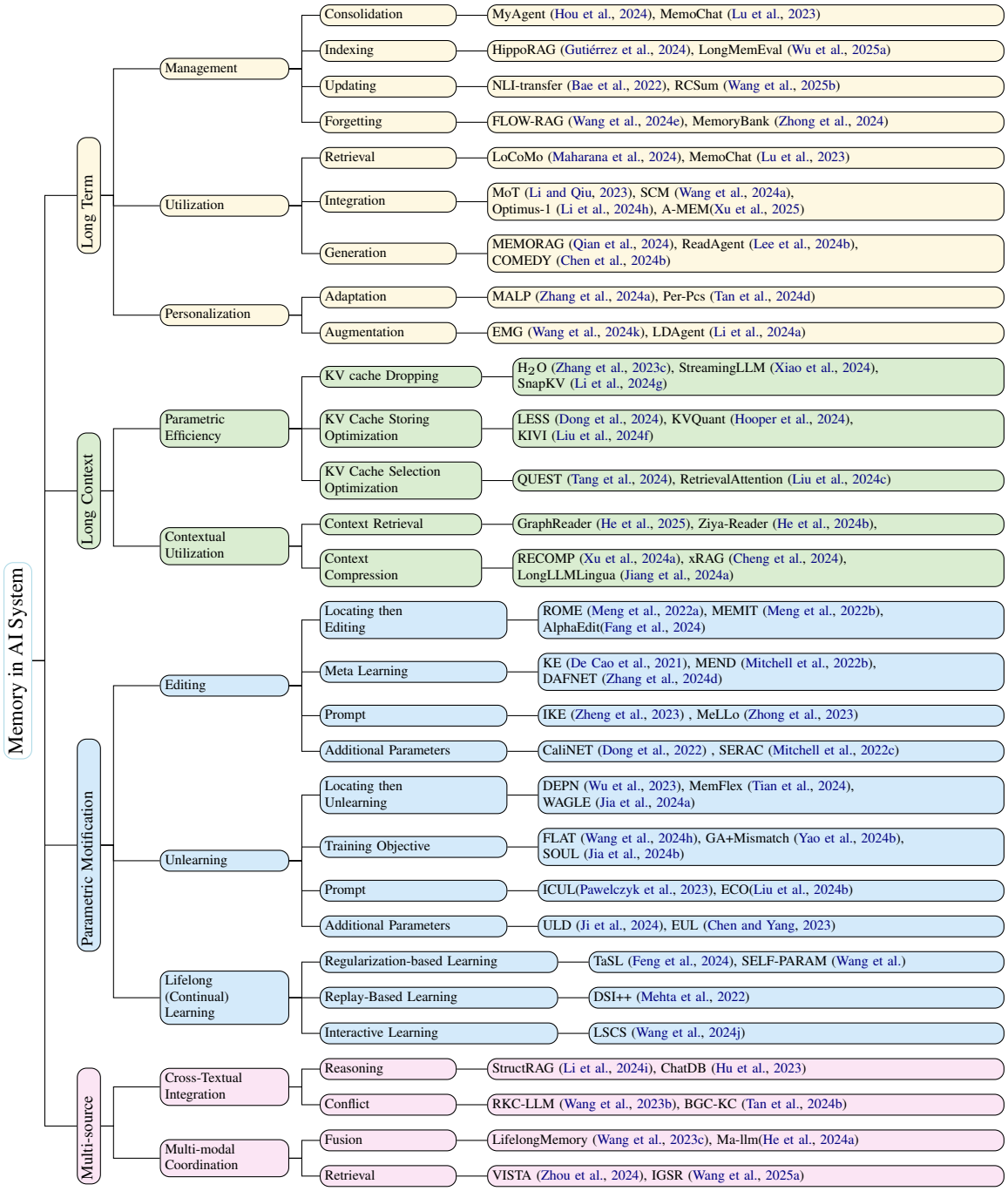

Figure 6: Operation-driven system-level topics in AI systems.

knowledge either in model parameters via training on unstructured text or transforming it into a fixed external memory format. Updating and forgetting are mainly associated with knowledge editing and unlearning, typically within parametric memory. These directions aim to incrementally modify parameters in the model based on external input. However, due to the opaque nature of model internals, such memory operations remain at an early stage of active exploration. In contrast, memory indexing mechanisms for LLMs have received limited attention.

From the topic perspective, parametric modification studies are mostly centered on parametric memory, though some works attempt parameter adaptation through continual learning over unstructured text. Research under the long-context theme primarily focuses on compression and retrieval within unstructured memory, with some leveraging parameterized forms like key-value caches. In long-term memory studies, the emphasis is also on unstructured memory, particularly in terms of con-

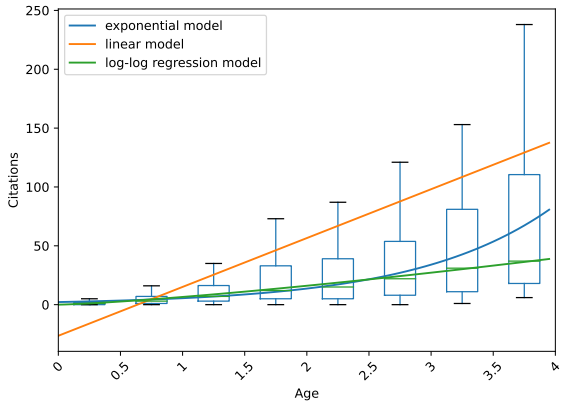

Figure 7: Boxplot of citation distributions from the 3,932 papers with respect to age, red curve is the expected citations $\hat{C}_i$. Generally $RCI >= 1$ indicate the paper is above median citations in its age group, and higher RCI indicate higher research impact.

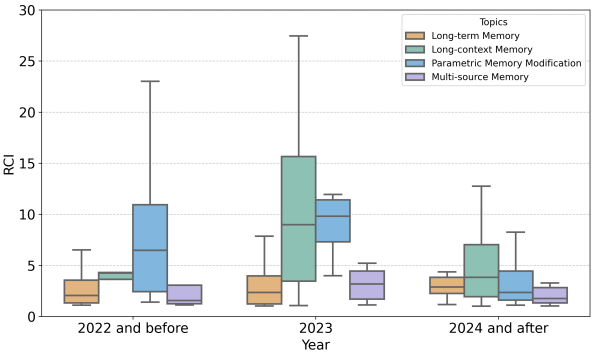

Figure 8: Overall distribution of median RCI across topics and years

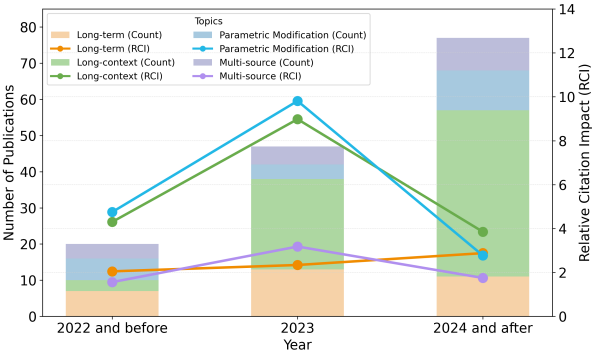

Figure 9: Overall temporal trends of topic-wise publication volume and median RCI.

solidation, compression, and retrieval. Research related to multi-source memory is still limited and typically involves integrating structured and unstructured information.

In summary, the limited exploration of contextual structured memory highlights an opportunity to develop more comprehensive memory operations by integrating it with unstructured memory. Second, research on multi-source memory remains scarce, despite the substantial challenges it poses—particularly the issue of memory conflicts arising from heterogeneous sources. Designing robust and consistent strategies for multi-source memory integration is thus a promising direction. Finally, although indexing has been extensively studied in traditional database systems, it remains underexplored in the context of LLM-based agents. The complexity of memory types and the need for vectorized or sparse retrieval methods call for new indexing approaches specifically tailored to reasoning and interaction in LLMs.

D.2 Memory Interactions Across Conference Venues

In addition to our primary paper collection, we also analyzed 81 method-focused papers with RCI $\geq$ 1 across major conferences. As shown in Figure 12, from the operation perspective, compression, forgetting, and updating appear more frequently in ML conferences (ICLR, ICML, NeurIPS), while retrieval and consolidation are more commonly featured in NLP conferences (ACL, EMNLP, NAACL). This distribution suggests that the former set of operations is still in the stage of theoretical exploration, whereas the latter is more grounded in practical application. Consequently, compression, forgetting, and updating still hold substantial potential for translation into real-world systems.

Indexing remains underrepresented in both ML and NLP venues. This may be partly due to its frequent co-occurrence with retrieval, and partly because current vector-based indexing approaches are relatively uniform, with few novel alternatives available.

From the topic perspective, long-term memory is more frequently addressed in NLP conferences, while long-context topics are more common in ML venues—likely reflecting the differing application- and theory-oriented focuses of these communities. Parameter modification appears more often in ML conferences, whereas multi-source memory is more prevalent in NLP conferences, highlighting the fact that multi-source memory challenges often arise during real-world applications and system integration.

Topic Name	Definition in Prompt
Long-Term Memory	**Definition:** Creating systems that ensure knowledge from past interactions remains accessible as new tasks emerge, maintaining continuity in multi-turn conversations. **Features:** Memory retention, retrieval, and attribution—preserving, accessing, and contextualizing memory to support coherent interaction.
Long-Context	**Definition:** Efficiently processing, interpreting, and utilizing very long input sequences without performance degradation. **Features:** Optimized attention, context compression, and mitigation of the "lost-in-the-middle" problem.
Parametric Memory Modification	**Definition:** Managing and updating internal parameters to preserve accuracy, privacy, and adaptability without full retraining. **Features:** Selective unlearning, precise model editing, distillation, and lifelong learning.
Multi-Source	**Definition:** Integrating and harmonizing diverse data types into a unified framework while resolving inconsistencies. **Features:** Multi-modal fusion, semantic consistency, conflict resolution, and redundancy removal.
Personalization*	**Definition:** Building user-centric memory systems that adapt to individual preferences and history while preserving privacy. **Features:** Privacy-aware profiling, consistent personalization, and long-term continuity.

Table 2: Definitions and features of the five memory-centric evaluation topics. *Personalization is treated as a specialized form of long-term memory that focuses on user-centric adaptation across sessions.

Prompts of the Relevance Evaluation to Task Definitions

System Instruction: Given the task and the abstract, evaluate the relevance of the abstract to the task.
Prompt Template:
"""
You are tasked with evaluating the relevance of a given article to a specific task definition.
Please read the following task definition, article title, and abstract carefully.
Based on the content, rate the relevance on a scale from 1 to 10,
where 1 means not relevant at all, and 10 means highly relevant.
Task Definition: $\{task_{def}\}$
Article Title: $\{title\}$
Abstract: $\{abstract\}$

Please provide your rating in the format [[Rating]].
For example, if the relevance is high, you might respond with [[9]]. """

Figure 10: Prompt for evaluating article relevance to specific task definitions.

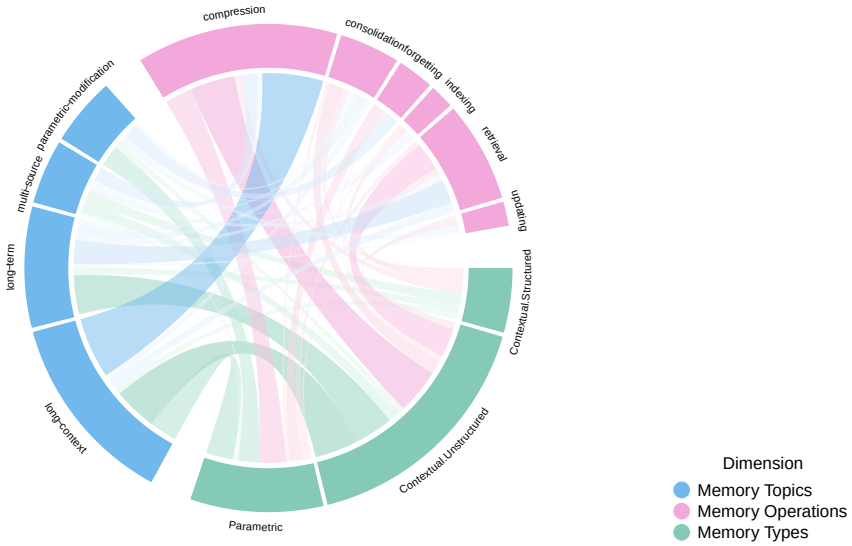

Figure 11: Chord Map of Interactions Across Memory Topics, Operations, and Types.

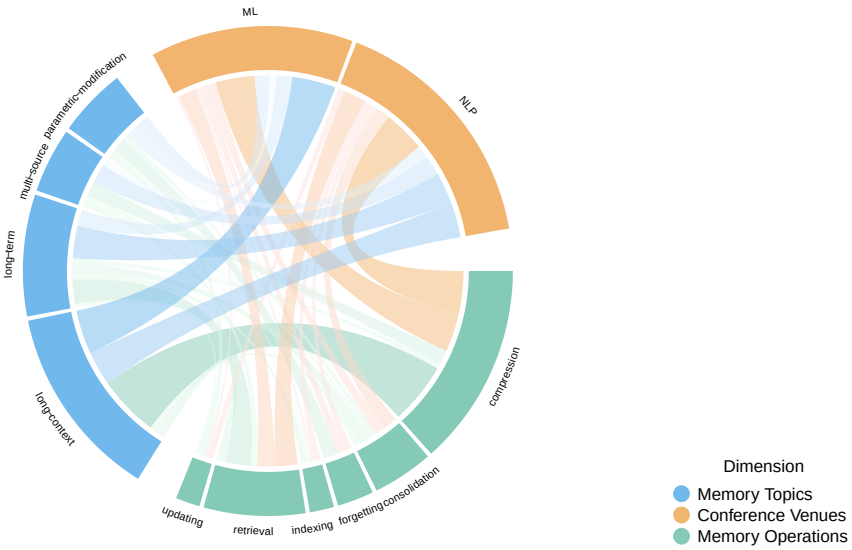

Figure 12: Chord Map of Interactions Across Memory Topics, Operations, and Conference Venues.

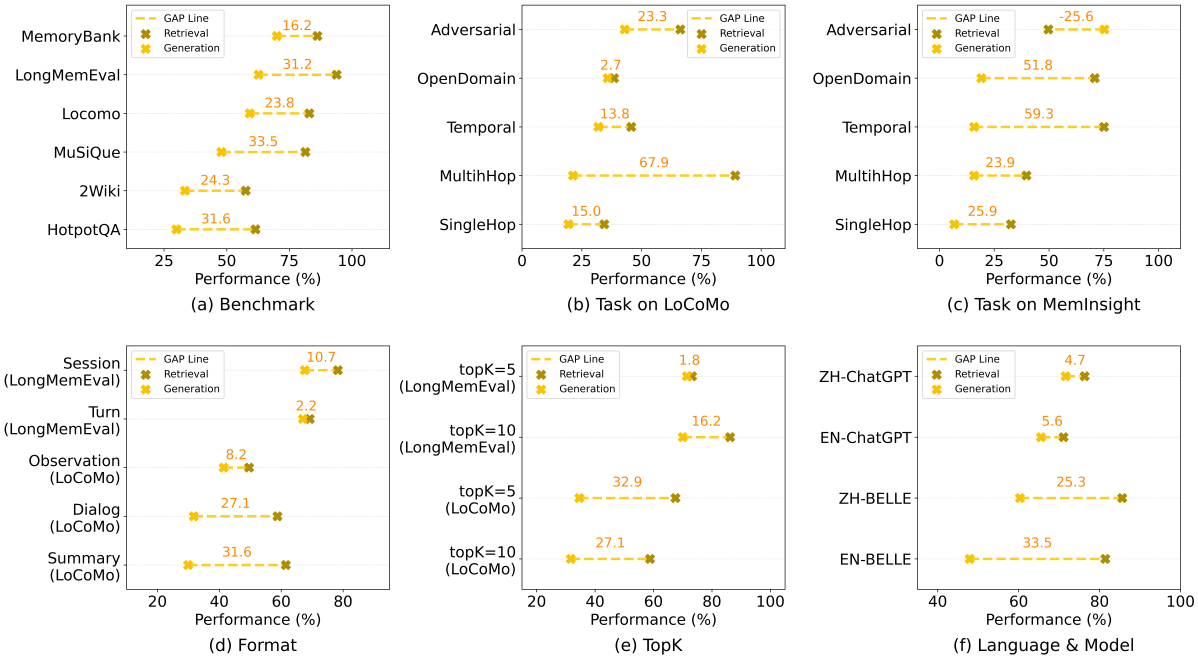

Figure 13: Retrieve-Generation GAP

Datasets	Mo	Operations	DS Type	Per	TR	Metrics	Purpose	Year	Access
LongMemEval (Wu et al., 2025a)	text	Indexing, Retrieval, Compression	MS	✗	✓	Recall@K, NDCG@K, Accuracy	Benchmark chat assistants on long-term memory abilities, including temporal reasoning.	2024	[LINK]
LoCoMo (Maharana et al., 2024)	text + image	Indexing, Retrieval, Compression	MS	✗	✓	Accuracy, ROUGE, Precision, 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	✓	✗	Accuracy, Human Eval	Enhance LLMs with long-term memory capabilities, adapting to user personalities and contexts.	2024	[LINK]
PerLTQA (Du et al., 2024)	text	Retrieval	MS	✓	✗	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., 2024)	text	Retrieval	MS	✓	✗	Accuracy	To evaluate dialogue systems under realistic, real-time, and long-context multi-party conversation conditions.	2024	[LINK]
CC (Jang et al., 2023)	text	Retrieval	MS	✗	✓	BLEU, ROUGE	For long-term dialogue modeling with time and relationship context.	2023	[LINK]
LAMP (Salemi et al., 2023)	text	Consolidation, Retrieval, Compression	MS	✓	✓	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 dialogue 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, Precision, PPL, BLEU, DISTINCT	For dynamic persona tracking and consistent long-term human-bot interaction.	2022	[LINK]
2WikiMultiHopQA (Ho et al., 2020)	table + knowledge 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	✗	✗	EM, F1	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 3: 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 DailyMail for evaluating long document summarization	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 summarization	2021	[LINK]
L-Eval (An et al., 2024a)	text	compression, retrieval	Rouge-L, F1, GPT4	Benchmark containing 20 sub-tasks specially 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]
SWE-bench (Jimenez et al., 2024)	text	compression, retrieval	Resolution rate (% Resolved)	Benchmarking LLMs' ability in solving GitHub issues. Consisting 2,294 task instances from 12 popular python repositories. Requiring LLMs to process very long context (reading the whole codebase with thousands of files).	2023	[LINK]
SWE-bench Multimodal (Yang et al., 2025)	text + image	compression, retrieval	Resolution rate (% Resolved), Inference cost (Avg. $ Cost)	Extending the original benchmark with image modal with 517 task instances.	2024	[LINK]
∞Bench (Zhang et al., 2024e)	text	compression, retrieval	F1, Acc, ROUGE-L-Sum	Benchmark containing 12 sub-tasks specially designed for evaluating extreme long context (on average surpassing 100K tokens) language models from different aspect.	2024	[LINK]
LooGLE (Li et al., 2024b)	text	compression, retrieval	Bleu-1, Bleu-4, Rouge-1, Rouge-4, Rouge-L, Meteor score, Bert score, GPT4 score	Benchmark containing 7 major tasks specially designed for evaluating extreme long context (each document surpass 24K tokens) language models from different aspect.	2023	[LINK]

Table 4: 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., 2024c)	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 5: 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]
MovieChat-1K (Song et al., 2024)	text + video	Retrieval	2	2	QA	Accuracy	For long-term video understanding for Large Multimodal Models across video question-answering and video captioning tasks.	2025	[LINK]
Context-conflicting (Tan et al., 2024b)	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, social 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 language queries over egocentric video with temporal memory.	2022	[LINK]
2WikiMultihopQA (Ho et al., 2020)	text	Indexing, Retrieval, Compression	2	1	Reasonin	EM, F1	Multi-hop QA requiring reasoning across two Wikipedia passages with sentence-level supporting evidence.	2020	[LINK]
HybridQA (Chen et al., 2021b)	text	Retrieval Compression	2	1	Reasonin	EM, F1	QA requiring reasoning across structured 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 snippets; often used as source for contradiction analysis.	2019	[LINK]
ComplexWebQuestions (Talmor and Berant, 2018)	text	Retrieval Compression	>1*	1	Reasonin	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, Supporting Fact Accuracy	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 disagreement analysis.	2017	[LINK]
WebQuestionsSP (Yih et al., 2016)	text	Indexing Retrieval Compression	>1*	1	Reasonin	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 6: 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	✓	✓	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., 2023b)	Augmentation	✗	✓	Retrieved & Context + Query	Response	Transformers	Retrieval	hierarchical transformer retriever, context-aware prefix encoder	2023	[LINK]
SiliconFriend (Zhong et al., 2024)	Augmentation	✗	✓	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	✗	✓	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., 2024d)	Adaption	✗	✗	User History	/	Llama-2-7B	Consolidation	modular PEFT sharing, collaborative personalization, user history assembly	2024	[LINK]
LAPDOG (Huang et al., 2023a)	Augmentation	✓	✓	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	✓	✓	Retrieved & Context + Query	Response	ChatGLM, BlenderBot, ChatGPT	Consolidation, Updating, Retrieval	long-term dialogue modeling, event & persona memory, mudular agent architecture	2025	[LINK]

Table 7: 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, memorization–retrieval–response cycle	2023	[LINK]
MemoryBank (Zhong et al., 2024)	Consolidation	✗	✓	✓	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	✓	✓	✓	Memory + Dialogue History	Response	T5	Consolidation, Updating, Retrieval	Session-level memory tracking, evolving dialogue system	2022	[LINK]
FLOW-RAG (Wang et al., 2024e)	Updating	✗	✓	✗	Knowledge Base + Query	Response	GPT4o, Gemini, llama2-7B-chat	forgetting	RAG-based unlearning	2024	[LINK]
FLARE (Jiang et al., 2023b)	Retrieval	✗	✓	✗	Database + Query	Response	WebGPT, WebCPM	retrieval	Active retrieval during generation, forward-looking query prediction	2023	[LINK]
HippoRAG (Gutiérrez et al., 2024)	Retrieval	✗	✓	✗	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	✗	✓	✓	Dialogue History + Query	Retrieved Results	Transformer++	Retrieval	Iterative query reformulation, context-aware query rewriting	2024	[LINK]
EWE (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	✗	✓	✗	Context + Query	Response	Mistral7B-Instruct, Phi-3-mini-128K-instruct, GPT-4o	Retrieval, Compression	Global memory retrieval, KV memory compression, Feedback-guided generation	2024	[LINK]
ReadAgent (Lee et al., 2024b)	Generation	✗	✓	✗	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	✗	✗	✗	Examples + Task Instruction	Trajectory + Thoughts	GPT4V, Qwen2VL	Updating	Trajectory abstraction memory, multi-modal, iterative reasoning correction	2025	[LINK]

Table 8: 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	Type	TF	DF	Operations	LMs	Features	Year	Code
StreamingLLM (Xiao et al., 2024)	KV Cache Dropping	✓	✗	Compression	Llama-2, MPT, PyThia, Falcon	Static KV cache dropping, Attention sink in the initial tokens	2024	[LINK]
FastGen (Ge et al., 2024)	KV Cache Dropping	✓	✗	Compression	Llama-1 7B/13B/30B/65B	Adaptive profiling-based KV cache dropping	2024	[LINK]
H_2O (Zhang et al., 2023c)	KV Cache Dropping	✓	✗	Compression	OPT, Llama-1, GPT-NeoX	Dynamica KV cache dropping, Retain Heavy Hitter tokens	2023	[LINK]
SnapKV (Li et al., 2024g)	KV Cache Dropping	✓	✗	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]
Scissorhands (Liu et al., 2023d)	KV Cache Dropping	✓	✗	Compression	OPT 6.7B, 13B, 30B, 66B	Dynamic KV cache dropping, Persistence of importance hypothesis	2023	[LINK]
FlexGen (Sheng et al., 2023)	KV Cache Storing Optimization	✓	✓	Compression	OPT 6.7B to 175B	KV cache quantization and offloading	2023	[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	✓	✓	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	✓	✓	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	✓	✓	Retrieval	LongChat-7B-v1.5-32K, Yarn-Llama2-7B-128K	Query-aware KV cache selection	2024	[LINK]
Memorizing Transformers (Wu et al., 2022a)	KV Cache Selection	✗	✓	Retrieval	Transformers	External KV cache memory	2022	[LINK*]
TokenSelect (Wu et al., 2025b)	KV Cache Selection	✓	✓	Retrieval	Qwen2 7B, Llama-3 8B, Yi-1.5-6B	Dynamic token-level KV cache selection	2025	[LINK]

Table 9: 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. [LINK]* indicates unofficial implementations.

Method	Type	SM	TM	Operations	LMs	Features	Year	Code
GraphReader (Li et al., 2024d)	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	2025	N/A
Ziya-Reader (He et al., 2024b)	Context Selection	T	T	Retrieval	Ziya2-13B-Base (LLaMA-2-13B)	Supervised finetuning, Position agnostic multi-step QA	2024	[LINK]
FILM (An et al., 2024b)	Context Selection	T	T	Retrieval	FILM-7B (Mistral 7B)	Data driven approach, lost in the middle	2024	[LINK]
xRAG (Cheng et al., 2024)	Context Compression	T	P	Compression	Mistral-7b and Mixtral-8x7b	Soft prompt compression	2024	[LINK]
AutoCompressor (Chevalier et al., 2023)	Context Compression	T	P	Compression	OPT-1.3B, 2.7B, LLaMA-2-7B	Soft prompt compression	2023	[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	2024	[LINK]
LongLLMLingua (Jiang et al., 2024a)	Context Compression	T	T	Compression	GPT-3.5-Turbo-06136, LongChat-13B-16k	Hard prompt compression	2024	[LINK]
LLMLingua-2 (Pan et al., 2024)	Context Compression	T	T	Compression	xlm-roberta-large, multilingual-BERT	Hard prompt compression, Data distillation	2024	[LINK]
QGC (Cao et al., 2024)	Context Compression	T	T	Compression	LongChat-13B16K, LLaMA-2-7B	Query-guided dynamic context compression	2024	[LINK]

Table 10: 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	✗	✓	✓	✓	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	✗	✓	✓	✗	aguila-7b	MEMAT is expanded upon MEMIT with **attention heads corrections** for **cross-lingual** editing.	2024	[LINK]
DEM (Huang et al., 2024b)	locating-then-editing	✗	✓	✓	✗	gpt-j-6b, llama2-7b	Use a **dynamic aware module** to select the editing layers. Evaluate **commonsense knowledge editing in free-text**.	2024	[LINK]
PMET (Li et al., 2024e)	locating-then-editing	✗	✓	✓	✗	gpt-j-6b, gpt-neox-20b	**Simultaneously optimize attention heads and FFN** but only update FFN weights.	2023	[LINK]
MEMIT (Meng et al., 2023)	locating-then-editing	✗	✓	✓	✗	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	✗	✓	✗	✗	gpt2-xl-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., 2024d)	meta learning	✗	✗	✗	✓	gpt-j-6b, llama2-7b	Supports **sequential** editing through **Intra-editing Attention Flow** (within facts) and **Inter-editing Attention Flow** (across facts).	2024	[LINK]
MALMEN (Tan et al., 2024a)	meta learning	✗	✗	✓	✗	bert-base, gpt-2, t5-xl, gpt-j-6b	Use least squares to merge edits reliably and decouple networks to save memory. Support **massive batch** editing.	2023	[LINK]
MEND (Mitchell et al., 2022a)	meta learning	✗	✗	✓	✗	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	✗	✗	✓	✗	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	✓	✓	-	-	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	✓	✓	-	-	vicuna-7b, gpt-j-6b	**Question Decompose + Self Check**	2023	[LINK]
Larimar (Das et al., 2024)	additional parameters	✓	✓	✓	✓	gpt2-xl, 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	✓	✗	✓	✓	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	✓	✗	✓	✓	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	✓	✗	✓	✗	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	✓	✗	✓	✓	t5-large, bert-base, blenderbot-90m	Scope Classifier + **Counterfactual Model**. Sequentially or simultaneously applying k edits yields the same edited model.	2022	[LINK]
GRACE (Mitchell et al., 2022c)	additional parameters	✓	✗	✗	✓	t5-small, bert-base gpt2-xl-1.5b	Support sequential editing by maintain **a codebook with a deferral mechanism** to decide whether to use the codebook for a input.	2022	[LINK]

Table 11: 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	Type	PR	TF	BUS	SUO	LMs	Main Advancement	Year	Code
ULD (Ji et al., 2024)	additional parameters	✓	✗	✓	✗	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	✓	✗	✓	✓	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	✓	✗	✓	✗	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	✓	✓	-	-	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	✗	✗	✓	✗	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]
DEPN (Wu et al., 2023)	locating-then-unlearning	✓	✓	✓	✗	bert-base	Detect and disable privacy-related neurons in language models to reduce data leakage.	2023	[LINK]
SOUL (Jia et al., 2024b)	training objective	✗	✗	✓	✓	opt-1.3b, llama2-7b	Unveil the power of **second-order optimizer** in LLM unlearning.	2024	[LINK]
SKU (Liu et al., 2024e)	training objective	✗	✗	✓	✓	opt-2.7b, llama2-7b, llama2-13b	Applies a two-stage framework combining harmful knowledge learning and task vector negation for effective unlearning.	2024	[LINK]
GA+Mismatch (Yao et al., 2024b)	training objective	✗	✗	✓	✗	opt-1.3b, opt-2.7b, llama2-7b	**Pioneered LLM unlearning** with an objective blending forgetting, random mismatch, and KL-based preservation.	2023	[LINK]
KGA (Wang et al., 2023a)	training objective	✗	✗	✓	✗	bart-base, distil-bert, lstm	Aligns knowledge gaps between models trained with retain vs. forget data to simulate forgetting via distributional divergence minimization.	2023	[LINK]

Table 12: 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
HippoRAG 2 (Gutiérrez et al., 2025)		✗	✗	Task-Free	Question Answering		Employs a training objective that **minimizes the Kullback-Leibler (KL) divergence** between the predictions of the original model and target model.	2025	[LINK]
SELF-PARAM (Wang et al.)	Regularization-based Learning	✓	✓	Task-Free	Question Answering	Llama-3.3-70B-Instruct	Enhances Personalized PageRank-based retrieval with deeper passage integration and online LLM usage, achieving superior performance on factual, associative, and sense-making memory tasks.	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	✗	✗	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	✗	✓	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	✗	✓	CIL	Object detection	LUCIR, PODNet	Enhances **memory replay by compressing data**, balancing sample quality and quantity for continual learning.	2022	[LINK]

Table 13: 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	Type	TF	STs	SNs	Input	Output	LMs	Ops	Features	Year	Code
GoG (Xu et al., 2024c)	reasoning	✓	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., 2023b)	conflict	✓	model + text	prompt + context	entities	answer	ChatGPT	Compression	Conflict span localization, instruction-guided conflict handling	2024	[LINK]
BGC-KC (Tan et al., 2024b)	conflict	✓	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 (?)	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]
DIVKNOWQ (Zhao et al., 2024b)	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., 2024i)	reasoning	✗	KG + Table + text	Loong, Podcast Transcripts	documents + query	answer	Qwen2-7B, 72B	Retrieval, Compression	Cognitive-inspired structurization, dynamic structure selection	2023	[LINK]

Table 14: 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	✓	✓	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	✓	✗	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	✓	✗	text + image	instruction + instances	response	OFA	compression	Cross-modal transfer learning	2023	[LINK]
NextChat (Zhang et al., 2023a)	fusion	✗	✓	text + image + boxes	image + text	response	CLIP	compression	Cross-modal alignment	2023	[LINK]
UniTranSeR (Ma et al., 2022)	fusion	✗	✓	text + image	context	response	MLM + MPM	compression	Intention-aware response generation, unified transformer space	2022	[LINK]

Table 15: 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 16: **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 17: **Framework-Level** Tools for Memory Management and Utilization.

Memory Tool	Level	Taxonomy	Operation	Function	Input/Output	Example Use	Source Type	Access
Mem0 (Taranjeet 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 18: **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 19: **Product-Level** Tools for Memory Utilization.

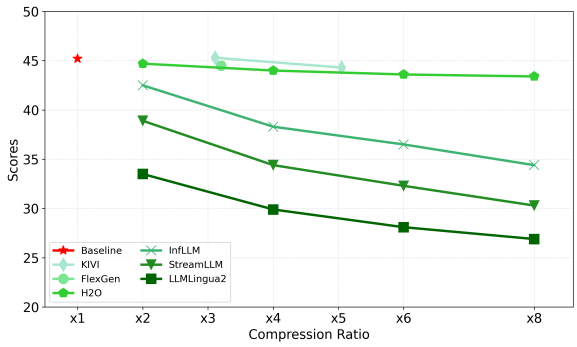

Figure 14: Compression based method performance with respect to compression rate on LongBench (Bai et al., 2024). Data borrowed from Yuan et al. (2024).

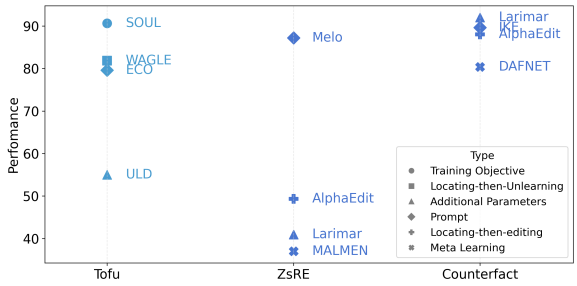

Figure 15: SOTA solutions across different categories on the CounterFact (editing), ZsRE (editing) and TOFU (unlearning) benchmark.

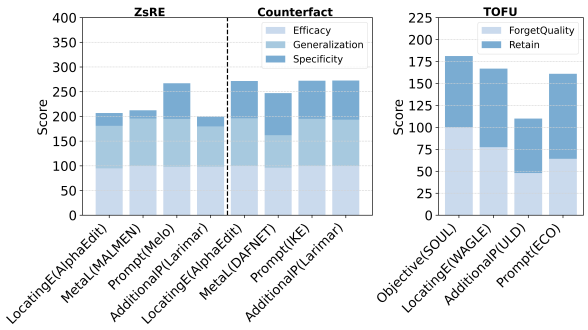

Figure 16: The sub-score distribution of SOTA solutions on the CounterFact (editing), ZsRE (editing) and TOFU (unlearning) benchmark.

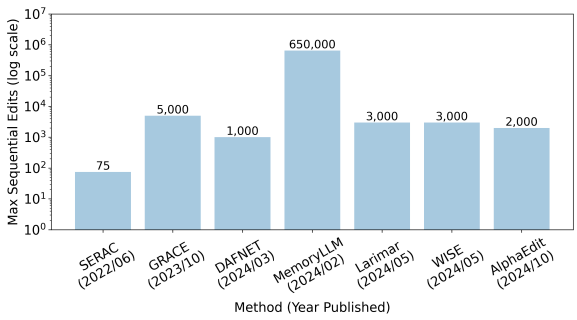

Figure 17: Maximum editing number of sequence editing in empirical experiments.

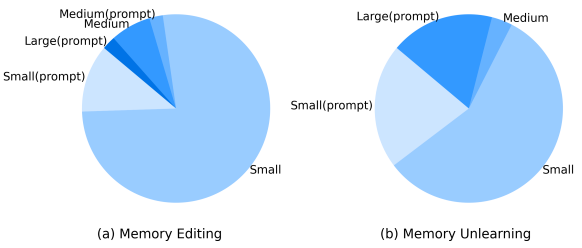

Figure 18: Model size distribution in memory editing and unlearning.

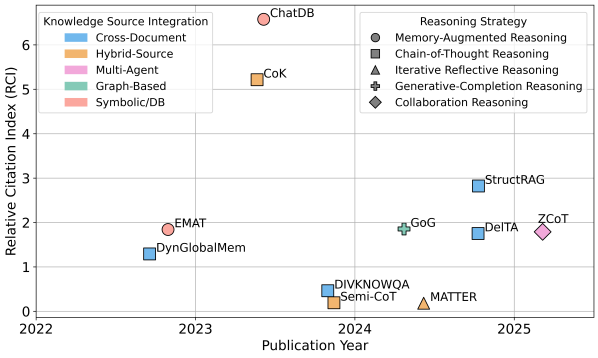

Figure 19: Trends in cross-textual reasoning: memory sources and reasoning strategies.

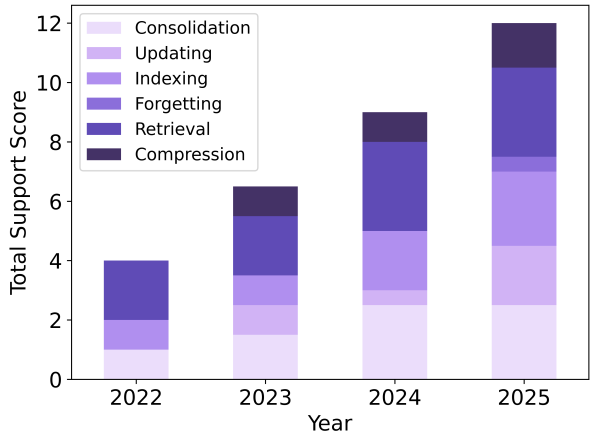

Figure 20: Evolution of memory operation support across Years.

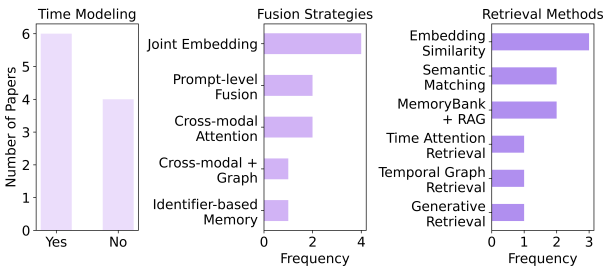

Figure 21: Analysis of temporal modeling, fusion strategies, and retrieval methods in multi-modal coordination.