Memory Meets (Multi-Modal) Large Language Models: A Comprehensive Survey

Anonymous authors
Paper under double-blind review

Abstract

Memory plays a foundational role in augmenting the reasoning, adaptability, and contextual fidelity of modern Large Language Models (LLMs) and Multi-Modal LLMs (MLLMs). As these models transition from static predictors to interactive systems capable of continual learning and personalized inference, the incorporation of memory mechanisms has emerged as a central theme in their architectural and functional evolution. This survey presents a comprehensive and structured synthesis of memory in LLMs and MLLMs, organizing the literature into a cohesive taxonomy comprising implicit, explicit, and agentic memory paradigms. Specifically, the survey delineates three primary memory frameworks. *Implicit memory* refers to the knowledge embedded within the internal parameters of pre-trained transformers, encompassing their capacity for memorization, associative retrieval, and contextual reasoning. Recent work has explored methods to interpret, manipulate, and reconfigure this latent memory. *Explicit memory* involves external storage and retrieval components designed to augment model outputs with dynamic, queryable knowledge representations—such as textual corpora, dense vectors, and graph-based structures—thereby enabling scalable and updatable interaction with information sources. *Agentic memory* introduces persistent, temporally extended memory structures within autonomous agents, facilitating long-term planning, self-consistency, and collaborative behavior in multi-agent systems, with relevance to embodied and interactive AI. Extending beyond text, the survey examines the integration of memory within multi-modal settings, where coherence across vision, language, audio, and action modalities is essential. Key architectural advances, benchmark tasks, and open challenges are discussed, including issues related to memory capacity, alignment, factual consistency, and cross-system interoperability. By charting the current landscape and identifying critical research directions, this survey aims to inform the development of memory-augmented (M)LLMs that are more flexible, context-sensitive, and aligned with the requirements of real-world intelligent systems.

Contents

"Memory is the treasury and guardian of all things."

— Marcus Tullius Cicero, De Oratore

1 Introduction

Recent advancements in artificial intelligence (AI) have led to the development of sophisticated systems, notably (Multi-Modal) Large Language Models (LLMs), which exhibit remarkable capabilities across various domains, from natural language processing and artificial intelligence to software engineering and social sciences (Brown et al., 2020; Hoffmann et al., 2022; Kaplan et al., 2020). Their proficiency extends to tasks such as multi-step reasoning (Wei et al., 2022; Li et al., 2025a) and cross-task generalization (Chatterjee et al., 2024), showcasing their potential to revolutionize various applications. The continuous development of LLMs is a crucial step towards achieving Artificial General Intelligence (AGI), necessitating the incorporation of advanced features that enable these models to autonomously explore and learn from real-world environments.

A pivotal aspect of this development is the integration of memory modules within LLMs, which play a fundamental role in how an agent accumulates knowledge, processes historical experiences, and retrieves information to inform decision-making and actions. By embedding memory capabilities, LLMs can evolve from static entities that rely solely on pre-trained knowledge to dynamic agents capable of continuous learning and adaptation. This transformation allows models to retain and leverage past interactions, thereby enhancing their performance in complex tasks that necessitate long-term planning and a deep contextual understanding. For example, a personal assistant agent with memory capabilities can remember user preferences and previous interactions, thereby delivering more personalized and contextually appropriate responses. Similarly, a trip-planning agent can track user itineraries and preferences, optimizing the efficiency and accuracy of its recommendations. In the healthcare sector, memory-enabled models can maintain comprehensive patient histories, leading to more precise diagnoses and tailored treatment plans. In educational settings, such models can monitor student progress and customize educational content to meet individual learning needs. Additionally, in customer service, memory-equipped agents can provide more efficient and personalized support, significantly enhancing user satisfaction and operational efficiency.

Memory is crucial for enabling LLMs and multimodal models to retain and utilize information over extended sequences, which is essential for tasks requiring context awareness and integration of diverse data types. We categorize memory in these models into three types: implicit memory (§2), embedded within the model's parameters; explicit memory (§3), involving external storage and retrieval; and agent memory, which maintains a persistent state across interactions. Each type plays a pivotal role in enhancing the model's ability to perform complex, context-dependent tasks.

This survey aims to provide a comprehensive overview of the current state of research on memory in LLMs, offering insights into their development, functionality, and impact on the performance of LLMs and multimodal LLMs (MLLMs).

Related Surveys of Memory Before the emergence of large language models (LLMs), Khosla et al. (2023) has explored memory-augmented neural networks. Their work investigated a range of network architectures, such as Hopfield Networks and Neural Turing Machines, and examined various types of memory, including sensory, short-term, and long-term memory. They also established connections between psychological theories of memory and their applications in AI, introducing architectures inspired by human memory systems.

Zhang et al. (2024d) presents a comprehensive survey on the memory mechanisms of LLM-based agents, systematically reviewing the design and evaluation of memory modules. Compared to our work, their focus is specifically on agents, particularly emphasizing historical and trajectory memory acquired through agent-environment interactions. He et al. (2024d) and Jiang et al. (2024a) focus on long-term memory in AI systems. Notably, Jiang et al. (2024a) proposes that AI equipped with long-term memory, capable of storing and managing real-world interactions, can achieve self-evolution.

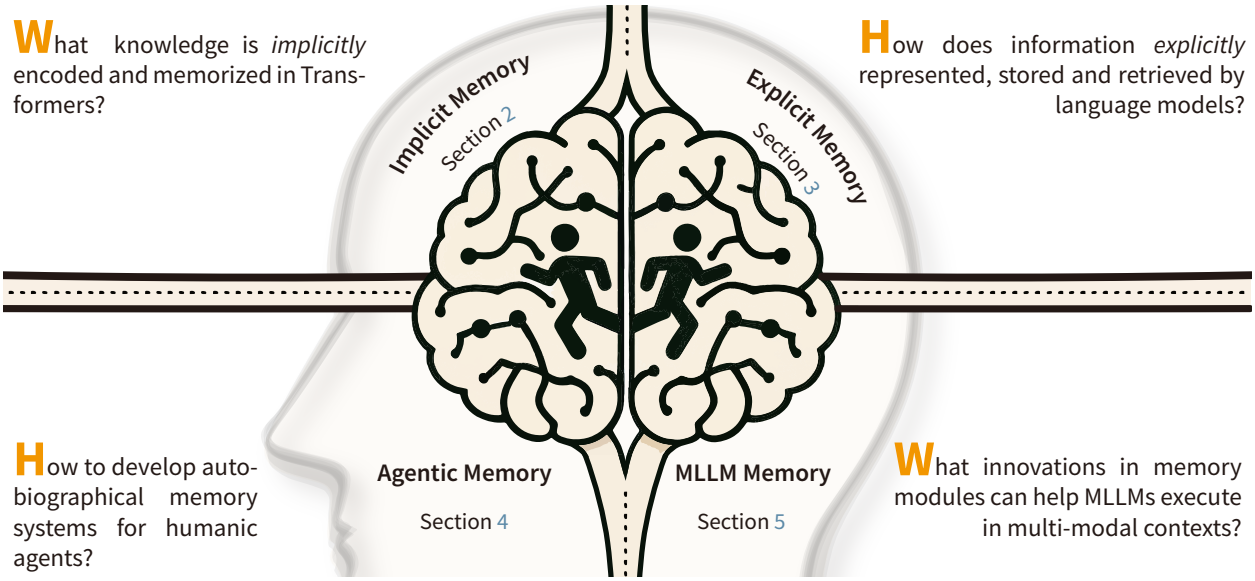

Figure 1: The overall framework of Memory Mechanisms in Large Language Models.

More recently, increasing attention has been devoted to memory in AI systems. Du et al. (2025) reconceptualize memory systems by categorizing them according to atomic operations and representation types, distinguishing between parametric and contextual forms of memory. They further classify memory operations into management and utilization, offering a detailed taxonomy and technical analysis that provides practical insights. Shan et al. (2025) explores the similarities and differences between human memory and memory in LLMs, discussing various forms such as text-based, KV cache-based, parameter-based, and hidden-state-based memory. Wu et al. (2025b) provides an in-depth analysis of memory in LLM-driven AI systems, categorizing memory-related methods across object, form, and temporal dimensions using an eight-quadrant framework.

While these surveys offer valuable perspectives, either drawing analogies with human memory or examining specific LLM-based memory forms and sources, none provide a unified view of memory study across pure LLMs, LLM-based agents, and further multimodal models. Consequently, our work presents a comprehensive survey that spans this full spectrum.

2 Implicit Memory: Unveiling Knowledge Inside Transformers

Implicit memory, a concept originating from psychology, refers to memories that are used unconsciously and are not stored explicitly Dew & Cabeza (2011). In the context of deep Transformer era, we define "implicit memory" as follows:

> **Implicit Memory** refers to the intrinsic information embedded within a model's parameters, encompassing self-knowledge, facts, commonsense, associative memory, and other related elements, which collectively enable the generation of contextually relevant responses across a variety of tasks.

The rise of Transformer models (Kovaleva et al., 2019; Brown, 2020; Geva et al., 2022b; Zhang et al., 2022; Stolfo et al., 2024) has brought significant attention to implicit memory due to their remarkable performance across multiple domains. Researchers are exploring how these models store and utilize knowledge within their parameters to understand and potentially enhance their capabilities. In this section, we attempt to answer the following research questions.

***RQ1**: What knowledge is implicitly encoded and memorized in Transformers?*
***RQ2**: How is information memorized, retrieved, and modified in Transformers?*

In the following subsections, we provide an overview of the current investigations on LLMs' memorization, including knowledge memorization and knowledge expression (§2.1.1), associative memory (§2.1.2), and implicit memory modification (§2.2). The structure of this section is demonstrated in Figure 2

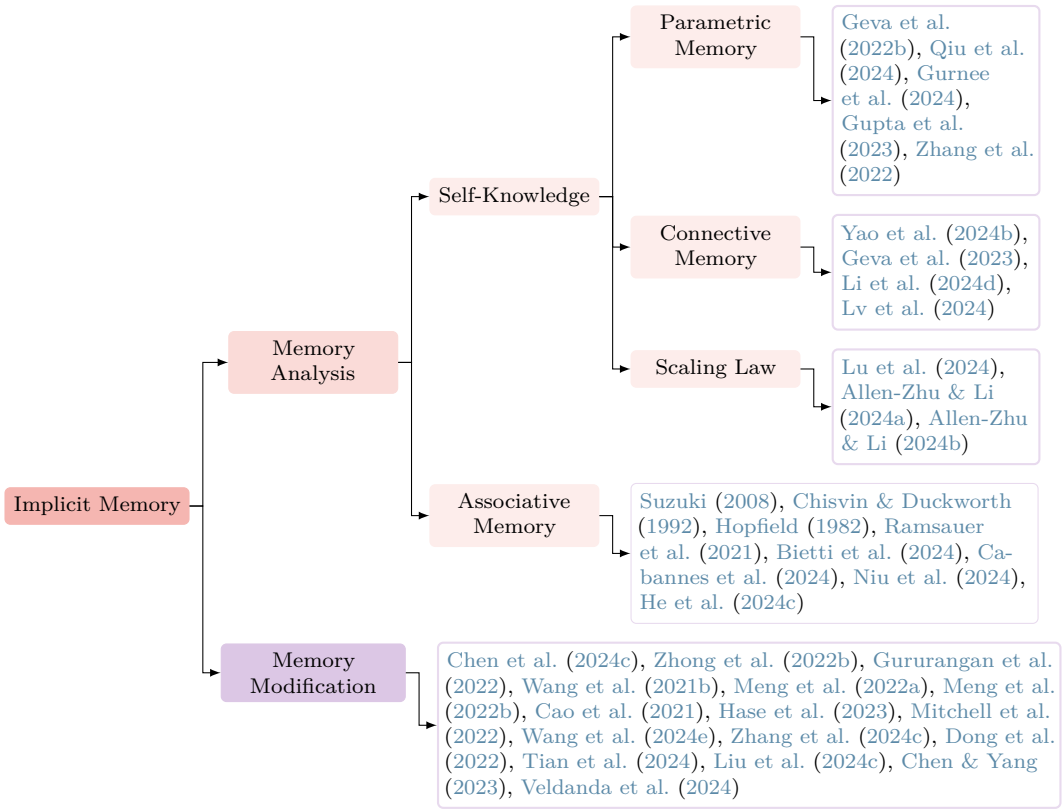

Figure 2: Taxonomy of implicit memory in Transformer.

2.1 Memory Analysis of Transformers

2.1.1 Knowledge Memorization

Transformers (Vaswani, 2017) have shown an impressive ability to memorize and retrieve knowledge implicitly stored in their parameters (Geva et al., 2023; Yao et al., 2024b; Lv et al., 2024; Stolfo et al., 2024). Studies have investigated how components like Feed Forward Networks (FFNs) and Self-Attentions (SAs) contribute to this knowledge memorization (Kovaleva et al., 2019; Geva et al., 2021; 2022b; Dai et al., 2021; Clark, 2019; Hoover et al., 2020; Yu et al., 2023a). This research is crucial for comprehending the internal workings of Transformer models and, accordingly, to manipulate their stored knowledge to enhance model performance.

Knowledge Memorized in Parameters There are two primary hypotheses (**H1&H2**) concerning the memorization of knowledge in transformer-style language models.

> **H1:** Knowledge is encoded through FFNs, emphasizing the role of feed-forward layers within transformer architectures in memorizing information.

An FFN is usually implemented as a stack of interleaved linear and non-linear layers. It has been used as a sub-module in many different neural architectures and has been shown to be crucial for Transformer's representation power (Geva et al., 2021; Meng et al., 2022a; Gupta et al., 2023). We categorize memory mechanisms within FFNs into two classes:

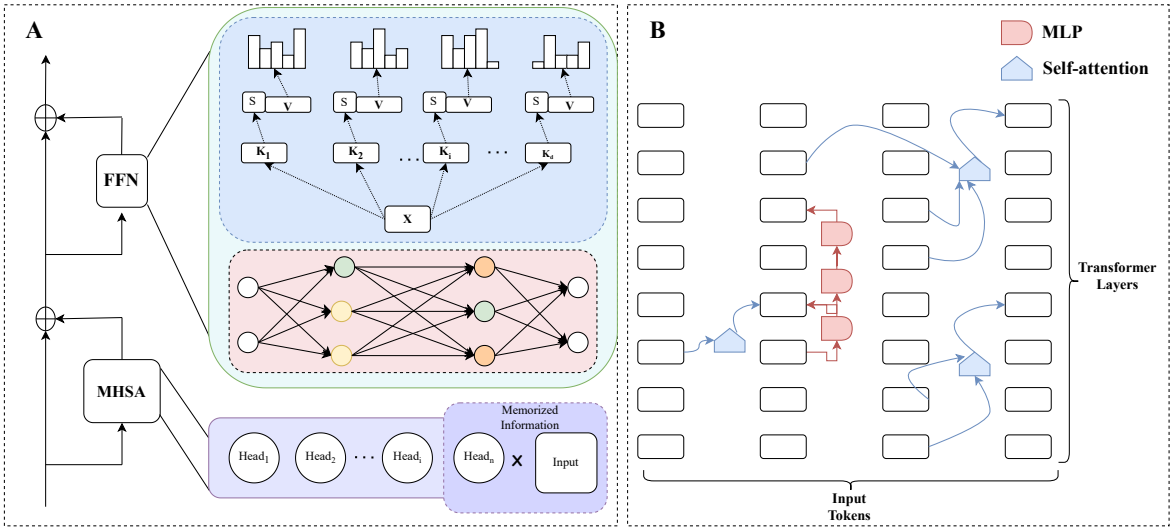

Figure 3: Parameter vs. Circuit. The left graph demonstrates the location of memory stored in Multi-Layer Perceptron layer and Self-Attention heads of the Transformer module, representing *FNNs act as key-value memories*, *Different FFN neurons memorize different information*, and *manipulating attention head distributions*, respectively. The right graph is a simplified demonstration of the knowledge flow between different transformer layers and various components within different layers.

FFNs act as key-value memories.[1] Geva et al. (2021) demonstrates that each key correlates with a specific set of human-interpretable textual patterns and each value induces a distribution over the output vocabulary. Based on this, Geva et al. (2022b) investigates the mechanism in which feed-forward layers update the inner representation, observing value vectors often encode human-interpretable concepts. Recently, Qiu et al. (2024) re-explores the key-value neural memories, conducting empirical ablation studies on updating keys or values in LLM. They recognize that updating the keys in a model is generally more effective than updating the values.

Different FFN neurons memorize different information. Dai et al. (2021) introduces the concept of knowledge neurons. They hypothesized that knowledge neurons in the FFN module are responsible for expressing facts. Zhang et al. (2022) shows that FFN neurons can be split into different functional partitions and some partitions are specialized in memorizing fact-related knowledge (Zhang et al., 2023f). Geva et al. (2022a) empirically finds that each vector of neurons in FFNs can be interpreted as a concept in the vocabulary space. Wu et al. (2023b) proposed a privacy neurons detector to locate neurons associated with private information. Stolfo et al. (2024) investigated entropy neurons and token frequency neurons which are two critical components believed to influence the representation and regulated uncertainty of LLMs.

H2: The attention mechanism is more crucial for knowledge storage, examining the relationship between the distribution of attention heads and the aggregation of knowledge.

Many works have analyzed Self-Attention layers for interpretability (Clark, 2019; Hoover et al., 2020), with *a focus on manipulating attention head distributions*. Recently, Yu et al. (2023a) study controlling LLMs to specifically leverage the in-context knowledge or facts memorized in pertaining by changing attention head distributions. Li et al. (2024c) identify a sparse set of attention heads that are completely related to fact knowledge of Alpaca Taori et al. (2023), and during inference, they shift activations along these truth-correlated directions, significantly improving the Alpaca truthfulness. Jiang et al. (2024c) further mathematically explores how Transformers can complete memory tasks based on the observation that LLMs'

[1]The feed-forward layer can be expressed as $\mathrm{FF}(\mathbf{x}) = f(\mathbf{x} \cdot K^\top) \cdot V$, where K denotes key vectors and V denotes value vectors, distinct from the key-value pairs used in self-attention mechanisms.

ability to retrieve facts can be easily manipulated by changing contexts. They theoretically prove and empirically show that the Transformer gathers information using self-attention.

Knowledge Flows in Connections As noted above, most studies have concentrated on knowledge storage within isolated components, such as FFNs and attention heads. In contrast, Yao et al. (2024b) studied connections between these Transformer components, introducing the concept of "knowledge circuits" to explore how different components collaborate to store and express knowledge. By ablating component-to-component connection, they demonstrate the knowledge circuit in LLMs for facts, linguistics, and commonsense. Previous work has explored knowledge flows similar to the knowledge circuit concept in LLMs. Geva et al. (2023) adopted the "knock out" strategy which blocks Multi-Layer Perceptron (MLP) or Multi-Head Self-Attention (MHSA) sublayers to investigate how the LLMs retrieved factual knowledge internally in inference. They conduct experiments on attribute prediction given subject-relation as queries, revealing two key components in the prediction process: the *subject enrichment process* where the early MLP sublayers are the primary source and the *attribute extraction operation* where the upper MHSA sublayers mainly carry out. Typically, Lv et al. (2024) explored several mechanisms employed by LLMs for factual recall tasks. They decompose MLP outputs into components that are easily understandable to humans based on linear regression, availably finding a universal anti-overconfidence mechanism in the final layer of models.

Scaling Law of Knowledge Memorization. The scaling law (Kaplan et al., 2020) has been used to describe model performance in terms of critical variables such as model size, dataset size, and the amount of computing used for training. *In the general pretraining field*, Kaplan et al. (2020) empirically study scaling laws for the performance of the language model on the cross-entropy loss L. They conclude an equation of scaling laws with model size and training time:

$$L(N, S_{\min}) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{S_c}{S_{\min}}\right)^{\alpha_S}, \tag{1}$$

where N represents non-embedding parameters, $S_{\min}$ represents the minimum number of steps necessary to reach L, $\alpha_N \sim 0.077$, $\alpha_S \sim 0.76$, $N_c \sim 6.5 \times 10^{13}$, and $S_c \sim 2.1 \times 10^3$.

For factual knowledge memorization, Lu et al. (2024) investigates the relationship between model size, training epochs, and fact memorization. Their study reveals that LLMs' fact knowledge capacity follows a linear and negative exponential relationship with model size and training epochs, respectively, suggesting that memorizing all public facts, like those in Wikidata[2], is nearly impossible in a general pre-training setting:

$$C = C^* - \alpha_E \cdot \exp(-\beta_E \cdot Epoch), \tag{2}$$

where C denotes fact capacity, C^* means the LLMs' fact capacity saturation when epochs approach infinity, and α_E and β_E are constants. Additionally, they find the scaling law of LLMs' fact memorization is similar to general pre-training, and the test loss L on fact generalization also follows the power-law (Kaplan et al., 2020), promising the generalization of unseen fact knowledge:

$$L(D) = D_c * D^{\alpha_D}, \tag{3}$$

where D is the number of training facts, D_c and α_D are constant numbers. Their study also analyzes the compatibility and preference of LLMs' fact memorization, highlighting its inefficiency in handling redundant facts and its preference for memorizing more frequent or difficult facts. Allen-Zhu & Li (2024a) also explores the knowledge capacity scaling laws of LLMs, providing a more accurate and flexible alternative to traditional methods which often rely on evaluating language models against real-world benchmarks. Different from Lu et al. (2024), they use a synthetic dataset rather than real-world facts to avoid benchmark contamination. Through comparison across different model architectures and types of knowledge, they conclude that a fully trained Transformer can store 2 bits of knowledge per parameter, even when quantized to int8, which is close to the theoretical maximum. Further, Allen-Zhu & Li (2024b) observe that mixed training with raw knowledge text and question-answer pairs yields better performance on out-of-distribution questions compared to the pretraining-finetuning approach on their synthesized biography dataset. They conclude that rewriting the pretraining data for knowledge augmentation and integrating more instruction-finetuning data during the pretraining stage can enhance LLM's knowledge memorization and extraction.

[2]https://www.wikidata.org/wiki/Wikidata:Main_Page

2.1.2 Associative Memory

In psychology, associative memory is the ability to build relationships between two previously unrelated features or ideas such as phone number and the name of a person (Suzuki, 2008). There are also other researchers trying to use associative memory in the physical computer memory architecture to overcome some basic problems of traditional address-based memory (Chisvin & Duckworth, 1992). In Transformer-based models, associative memory refers to the memory of two previously unrelated representations (*e.g.*, input-output vectors) learned during the training process.

Energy-based Model Mimic Associative Memory Hopfield Network (Hopfield, 1982) is a type of fullly-connected recurrent energy-based neural network that is used primarily for associative memory. This network structure leverages a set of interconnected neurons and weight matrices to encode multiple patterns, where each pattern is associated with a specific input. The storage of these patterns is facilitated by the modification of synaptic weights between neurons, a process that is governed by the principles of Hebbian learning. In details, the original Hopfield Network is composed with binary neurons $V_i \in \{0, 1\}$, so the instantaneous state of the system is defined by the combination of neurons' states $V_1, ..., V_n$. The initial value of the neuron states are all 0 as they are not "firing up" (Hopfield, 1982). The strength of one-way connection from neuron V_j and neuron V_i is denoted as T_{ij} where T_{ij} is computed using equation

$$T_{ij} = (2V_i - 1)(2V_j - 1) \; where \; T_{ii} = 0, \tag{4}$$

and the Non-connected neurons have strength $T_{ij} = 0$. The associative memory is stored in the format of patterns of states, and the connection strength T_{ij} makes up the weight matrix T. The state of the system changes based on a pre-defined step-function at a given time t. The changes to a single neuron V_i follows the step-function given the weight matrix elements T_{ij} and all the neurons that connects to the neuron V_i

$$V_i = \begin{cases} 1 & \text{if } \sum_{j \neq i} T_{ij} V_j(t) > U_i \\ 0 & \text{if } \sum_{j \neq i} T_{ij} V_j(t) < U_i \end{cases} \tag{5}$$

where U_i is a predefined threshold. Recent research on the Hopfield network (Ramsauer et al., 2021) has successfully integrated Dense Associative Memorys (DAMs) into modern deep learning architectures. This study introduces innovative updating rules and energy functions, transitioning the traditional discrete Hopfield network into a continuous framework. A key limitation of the original Hopfield network is its slow energy reduction during the pattern memorization phase. By utilizing a rectified polynomial energy function, DAMs accelerates energy decay, enabling the storage of more memory patterns within the same configuration space.

Transformer-based Models Recent literature suggests that the Transformer architecture stores associative memories in the form of outer products of finite-dimensional embeddings within the intermediate weight matrices. Bietti et al. (2024) analyzes the transformer structures with an associative memory point of view. The authors construct a synthetic bi-gram dataset and a two-layer, single-attention-head Transformer model. Through empirical analysis, the results demonstrate that the Transformer architecture employs the weight matrix at the output of the attention block as a repository for associative memories. These associations enable the key-query matrices to direct attention to relevant tokens, thereby enhancing the model's inductive capabilities. The experiments also show how associations are learned through training dynamics and can be used to remap input to output vectors. Jiang et al. (2024b) explores how LLMs use tokens within a given context as memory clues to retrieve memory patterns from their parameters, and how context can be leveraged to influence or "hijack" the output of LLMs.

Scaling Law of Associative Memory Cabannes et al. (2024) investigates the scaling law of the associative memory error rate in relation to model size and the number of data inputs, using a simple Transformer-based model. The error rate of the model is bounded by the previously seen data and the model capacity.

$$\mathcal{E}(f_q) \sim d^{-\alpha+1} + T^{-1+\frac{1}{\alpha}}, \tag{6}$$

where d is the model capacity, and T is the number of training samples. The model is denoted as f_q. α is the hyper-parameter that represent the Zipf exponent of the assumed data distribution. The errors in

this equation are measured by the recall accuracy of previously presented factual inputs. According to the equation, if the model possesses infinite memory, the error rate is constrained by the sample input size, as the model must learn sufficient associations to generalize the distribution. However, when memory is finite and the dataset size exceeds the memory capacity, the model attempts to remap new data into its memory space, which can result in memory interference when two distinct input-output relations are mapped to the same location. The paper also discusses various memory schemes for storing associations. Niu et al. 2024 conducted an empirical study on associative memory within the Transformer architecture and proposed a new energy function that, without introducing additional regularization terms, corresponds to a nearest-neighbor search over the memorized patterns.

Usage of Associative Memory Numerous studies have investigated the use of associative memory as a mechanism for storing patterns linked to past data points. For instance, CAMELoT (He et al., 2024c) introduces an additional module within the original attention layer of a Transformer model to enhance its ability to handle longer context windows. The authors suggest that compressing past context or inputs into associative memory conserves memory by eliminating redundancy, thereby freeing up space for new content in fresh memory slots. This approach also facilitates the replacement of outdated memory slots with more recent inputs. Additionally, the authors show how modifications to the energy function and update rule enable the modern Hopfield network to approximate the architecture of Transformer models. Krotov (2021) proposed an extension of the modern Hopfield network by incorporating an arbitrarily large number of recurrent layers, designed with a custom multi-layer recurrent model structure. Similarly, Millidge et al. (2022) proposed a generalized Hopfield network that decomposes a series of models—such as DAMs, Hopfield networks, and sparse distributed memories—into a three-stage framework comprising similarity, separation, and projection. This paper demonstrates that these three components not only generalize existing models but also extend their functional capabilities.

2.2 Implicit Memory Modification

Modifying implicit memory (Wang et al., 2024c) in language models involves altering the knowledge embedded within a model's parameters to enhance performance, reduce harmful outputs, and enable more efficient adaptation to new tasks. Focusing on methods for updating or removing knowledge without incurring the high costs of full retraining, as discussed in (Dai et al., 2021), we categorize memory modification into three main categories: **incremental training, memory editing, and memory unlearning**, as illustrated in Figure 4. **Incremental training** represents adding new knowledge into LLMs while building on the pre-existing information within them. **Memory editing** adjusts the embedded knowledge by modifying specific memory representations within LLMs. And **memory unlearning** eliminates incorrect or harmful internal knowledge from LLMs, thereby enhancing their reliability and trustworthiness. These approaches are intended to create more adaptable and efficient models, capable of rapidly incorporating new information while retaining consistent performance across various tasks. Techniques like dynamic updates, hyper-networks, and targeted unlearning play essential roles in enhancing the adaptability and dependability of LLMs for real-world applications.

2.2.1 Modification Methods

Incremental Training Incremental training in LLMs involves not only adding new knowledge but also ensuring consistency and alignment with the model's existing knowledge base. This process typically follows two main approaches: directly modifying the original parameters of the LLMs and adding knowledge through adapters, which store new memory separately.

Directly modifying the original parameters. Zhu et al. (2020) proposes a constrained fine-tuning method that selectively updates a subset of parameters to integrate new information efficiently. Expanding on Zhu et al. (2020), Padmanabhan et al. (2023) introduces context distillation to update knowledge using entity-specific texts while preserving the original LLMs distribution measured by KL-divergence. TRIME (Zhong et al., 2022b) incorporates in-batch examples as accessible memory during training. It improves performance by effectively leveraging local, long-term, and external memory with minimal computational overhead, achieving significant reductions in perplexity across various benchmarks. Another method, RECKONING (Chen et al.,

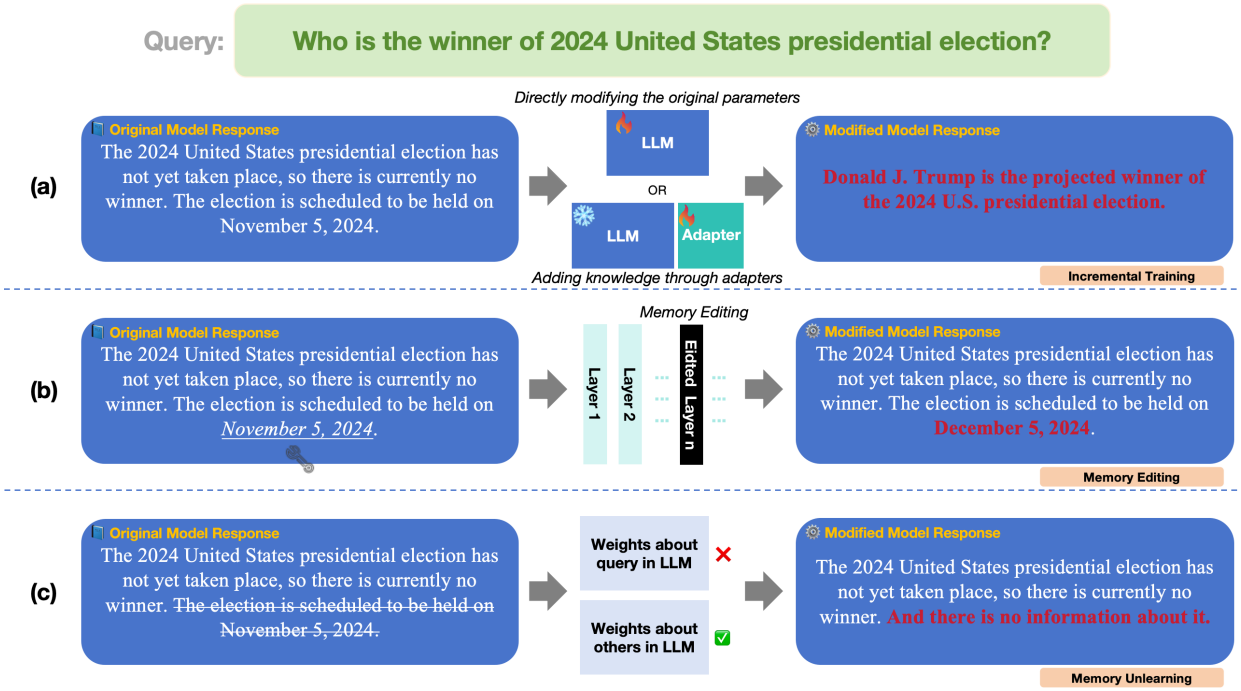

Figure 4: Three categories of Implicit Memory Modification

2024c) encodes context-based knowledge into model parameters via a bi-level learning process, which consists of two loops: the inner loop adapts the model to memorize facts, while the outer loop uses the updated parameters to answer reasoning questions.

Adding knowledge through adapters. Adapters provide a flexible means of incorporating new knowledge by leaving the original model parameters untouched. LoRA (Hu et al., 2022) uses low-rank decomposition matrices for targeted updates while maintaining the model's core structure. K-ADAPTER (Wang et al., 2021b) adds separate adapters for different knowledge types in models like RoBERTa. DEMiX (Gururangan et al., 2022) introduces domain-specific expert networks, training only the relevant expert for new knowledge. These methods balance memory augmentation and model stability, making language models more adaptable and resource-efficient.

Memory Editing The objective of knowledge editing is to incorporate new facts into a language model $\mathcal{M}_\theta$ through query-response pairs $\mathcal{D}_e = \{(q_i, x_i^*)\}_{i\in[1,N]}$. In this setup, q is the query that triggers the retrieval of factual knowledge from $\mathcal{M}_\theta$, such as "The president of the US is", and x is the intended response after editing, e.g., "Joe Biden". This integration is typically achieved by maximizing the probability of generating x^* based on q, which can be expressed as:

$$\max_\theta p_\theta(x^*|q) \quad \text{where} \quad (q, x^*) \sim \mathcal{D}_e \tag{7}$$

Certain knowledge editing methods first identify the implicit memory within large language models (LLMs) that corresponds to the targeted knowledge, and then modify the knowledge stored in the model's weights. ROME (Meng et al., 2022a) and MEMIT (Meng et al., 2022b) use the causal trace method to identify the regions in LLMs where relevant knowledge about triplet relations is stored. These methods treat the middle-layer MLPs of GPT models as associative memory structures with key-value associations, and modify the weights of these layers to insert new associations. Dong et al. (2022) propose CKA, a method to detect incorrect knowledge in PLMs. CKA compares the model's scoring of correct facts against fake facts to determine if the model has accurately learned the facts. It also introduces a method called CALINET, which adds new parameters to the model while keeping the original parameters fixed, allowing it to learn the

correct facts without overwriting other knowledge. KE (Cao et al., 2021) edits knowledge by employing a hyper-network to update the parameters of origin LLMs, so that LLMs can predict expected predictions for different inputs without affecting the prediction of any other input. Similarly, SLAG (Hase et al., 2023) utilizes a learnable hyper-network. This hyper-network takes model gradients as input and outputs new updates to be applied to the model parameters. Recently, the LAW (Wang et al., 2024e) method modifies specific MLP layer weights in a language model by adjusting the internal "key" and "values" associated with targeted knowledge, effectively disrupting the model's representation of that knowledge while preserving its reasoning abilities. FT-M (Zhang et al., 2024c) fine-tunes specific layers of the feed-forward network to maximize the probability of all tokens in the target sequence.

Memory Unlearning Knowledge unlearning can be described as follows: Given a training set $D = \{(x, y)\}$ for the language model $\mathcal{M}_\theta$, where x represents the input and y represents the corresponding label, we define D_f as the set of harmful and dangerous knowledge that we aim to forget, and D_r as the set of data we wish to retain. The goal of knowledge unlearning is to enable $\mathcal{M}_\theta$ to remove all information from D_f while preserving performance on D_r, which means:

$$\max_\theta dist\left(\mathcal{M}_\theta(D_f); \mathcal{M}'_\theta(D_f)\right) \quad \text{and} \quad \min_\theta dist\left(\mathcal{M}_\theta(D_r); \mathcal{M}'_\theta(D_r)\right) \tag{8}$$

Knowledge unlearning methods in LLMs can be categorized into distinct approaches based on their primary goals and techniques. Some methods, like a benchmark KnowUnDo (Tian et al., 2024) and a proposed method MemFlex (Tian et al., 2024) focus on precision unlearning, using targeted gradient manipulation to selectively remove sensitive information while preserving essential knowledge. In contrast, frameworks like SKU (Liu et al., 2024c) employ a two-stage approach, acquiring and systematically negating harmful knowledge to ensure safe responses while maintaining performance on benign tasks. Additionally, efficiency-focused methods, the Surgery framework (Veldanda et al., 2024) efficiently updates LLMs by unlearning outdated knowledge, integrating new information, and retaining performance on unchanged tasks using a three-part objective: reverse gradient for unlearning, gradient descent for updating, and KL divergence minimization for consistency. Lastly, solutions like EUL (Chen & Yang, 2023) prioritize selective and scalable unlearning by employing lightweight layers, facilitating iterative knowledge removal without impacting the model's overall capabilities, suitable for repeated applications where selective knowledge removal is essential.

2.2.2 Modification Benchmark

Memory Editing Benchmark Many memory editing benchmarks(Wang et al., 2023d; Cohen et al., 2024; Khandelwal et al., 2024) primarily focus on editing accuracy by constructing datasets designed to evaluate whether models can produce counterfactual responses when queried about specific factual knowledge. KnowEdit(Zhang et al., 2024c), for instance, includes a suite of six datasets tailored for assessing various knowledge editing methods. These datasets cover a diverse range of editing types, such as fact manipulation, sentiment alteration, and hallucination generation. This benchmark consolidates key evaluation criteria into four categories: edit success, portability, locality, and fluency, thereby providing a comprehensive evaluation framework for different editing approaches. MQuAKE(Zhong et al., 2023) offers a distinct perspective by evaluating whether edited models can answer multi-hop questions where the answer should logically change as an entailed consequence, revealing the limitations of prior methods for such questions. Eva-KELLM(Wu et al., 2023a) expands the scope of knowledge editing evaluation to a more general scenario where raw documents within datasets are directly utilized for editing. This benchmark offers greater generality and practical relevance, allowing for the assessment of various knowledge editing methods' performance in a multilingual context. The comparison of these three benchmarks can be seen in table 1.

2.3 Limitations, open questions, discussion

Current research in implicit memory often faces several key challenges:

- Generalization of findings: A significant limitation of many studies is that they focus primarily on knowledge memorization and extraction within the confines of specific tasks (e.g., relation triples prediction) or

Table 1: Comparison of Memory Editing Benchmarks Based on Different Dimensions.

Dimension	KnowEdit(Zhang et al., 2024c)	MQuAKE(Zhong et al., 2023)	Eva-KELLM(Wu et al., 2023a)
knowledge insertion, modification, and erasure	✓	✗	✗
Counterfactual and temporal updates	✗	✓	✓
Supports multilingual data	✗	✗	✓
High-quality, validated datasets	✓	✓	✓
Multi-task editing capabilities	✓	✗	✓
Includes reasoning tasks	✗	✓	✓
Emphasizes cross-lingual performance	✗	✗	✓
Handles multi-hop reasoning	✗	✓	✓
Suitable for dynamic system updates	✓	✓	✓
Efficient evaluation time	✓	✗	✓

particular types of knowledge (such as facts, commonsense, or bias-related knowledge). This narrow focus does not establish the generalizability or broader applicability of the conclusions drawn.

- Efficiency of probing methods: Research on knowledge circuit exploration is often hindered by the time-consuming nature of conducting numerous component-to-component ablations. This complexity can significantly slow down the process of systematically investigating model behavior and knowledge extractions.
- Risk of knowledge unlearning: While knowledge unlearning aims to remove unwanted or harmful information, it carries the potential risk of inadvertently disrupting related knowledge. This disruption can lead to unintended consequences and degraded performance in downstream tasks. Consequently, more comprehensive evaluations of models subjected to knowledge unlearning are needed to fully understand these effects.

Looking ahead, future research should focus on gaining a deeper understanding of the internal mechanisms of Transformer models and explore the development of more effective and efficient computational frameworks for implicit memory modeling.

3 Explicit Memory: When (M)LLMs Meet Retrieval

> In this work, we refer to **explicit memory** as specific, structured, or unstructured representations that store factual knowledge, history trajectories in an external storage.

Explicit memory allows the retriever to dynamically capture context-aware information and enables the generator to adaptively incorporate knowledge from external memory, thus improving the quality of generated outputs. It facilitates the retention and retrieval of information across sessions, ensuring continuity and enhancing the model's capacity to handle long contexts without exceeding its input limitations. Explicit memory plays a crucial role in providing flexibility and enhancing interpretability, especially in interactive, knowledge-intensive, and rapidly evolving domains.

In this section, we focus on the research question:

How is the explicit memory represented and utilized in different training and application scenarios?

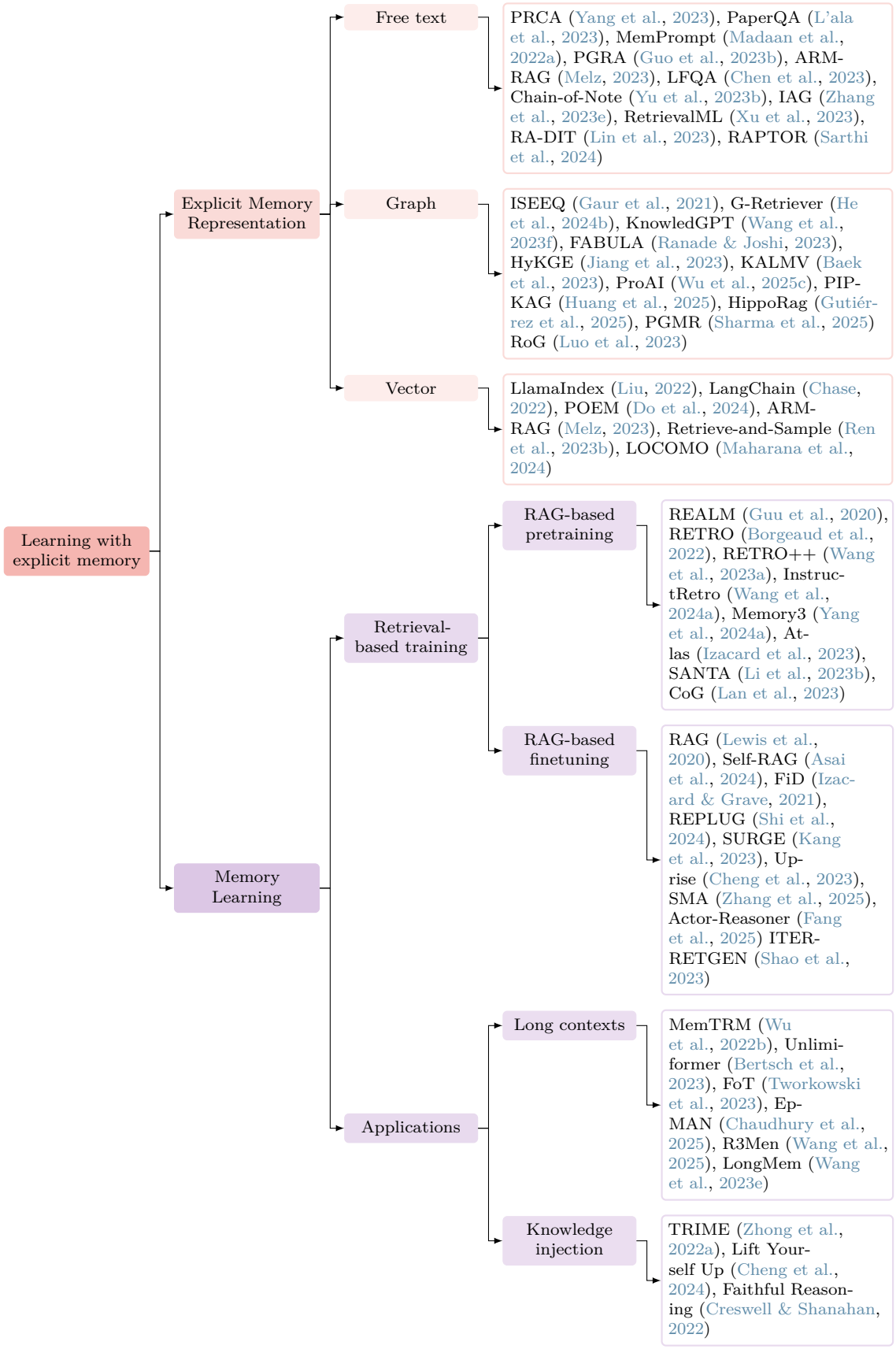

Figure 5: Taxonomy of the structure design for learning with explicit memory.

We will discuss how different types of memory are **externally represented and stored** (§3.1) the **interaction and updating of external knowledge during training** (§3.2), and how to **externalize implicit knowledge for retrieval** in typical scenarios (§3.3).

3.1 Explicit Memory Representation

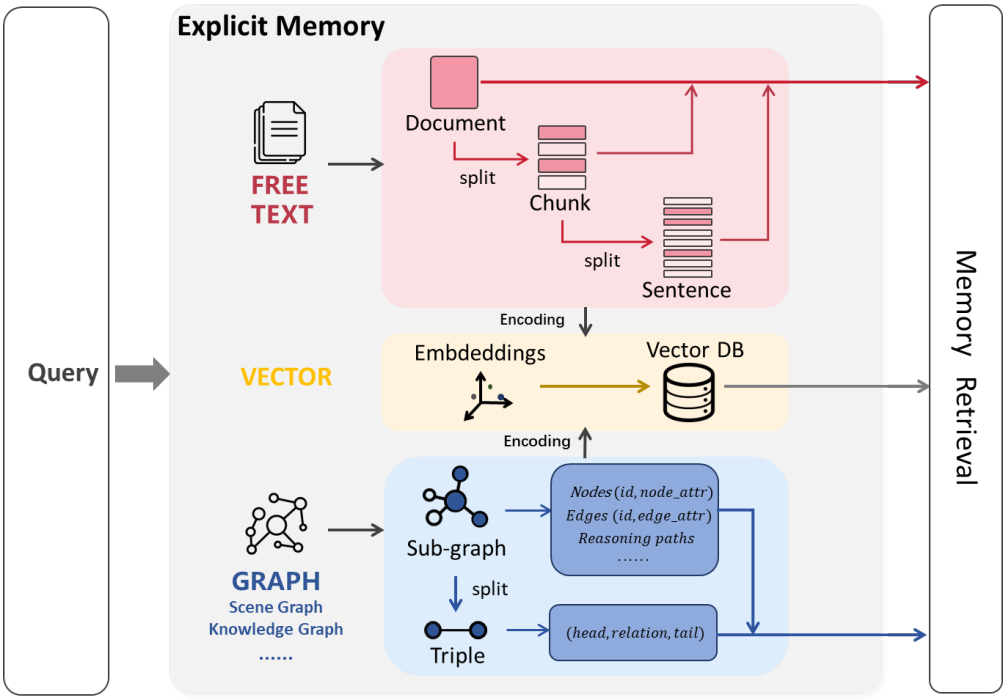

Figure 6: Three types of representations for explicit memory retrieval

3.1.1 Free text

LLMs are usually trained on free-form text, depending on target tasks. The free-form text is represented and stored at different levels of granularity as shown in Fig. 6:

Document As the largest memory unit, a document retains rich contextual information, which is crucial for understanding the overarching themes (Ren et al., 2023a). Commonly employed retrieval methods can be broadly categorized into sparse retriever and dense retriever. Sparse retrieval, exemplified by traditional search engines and the BM25 algorithm, relies on exact matches and TF-IDF weighting to retrieve lexically relevant documents (Asai et al., 2024; Ma et al., 2023). In contrast, dense retrieval focuses on semantic similarity (Shi et al., 2024; Melz, 2023) and is typically implemented using tools such as FAISS (Johnson et al., 2021) and pre-trained encoders. Once retrieved, this supplementary knowledge is concatenated with the question as input to the model. However, overly-long texts can introduce noise (Yang et al., 2023; Chen et al., 2023; Cuconasu et al., 2024; Zhang et al., 2023e), potentially obscuring specific facts (Trivedi et al., 2022; Yu et al., 2023b). Moreover, using entire documents as input increases inference time and computational costs for LLMs. Due to limited context window size in LLMs, excessively long texts will be truncated, risking the omission of critical data.

Chunk Dividing text into fixed-size chunks effectively reduces redundancy and strengthens the connection between retrieved memory and the query (Wang et al., 2024a; 2023a; Lin et al., 2023). There can be multiple chunks relevant to the query, but using all of them is usually infeasible due to the context length limit. To address this, RetrievalML (Xu et al., 2023) selects the top k chunks based on relevance, determining the optimal k value through experimentation. In contrast, PaperQA (L'ala et al., 2023) employs LLM-generated

relevance scores between the chunks and the query for more refined filtering. To preserve textual coherence, RAPTOR (Sarthi et al., 2024) clusters the chunks and generates summaries for each cluster. While chunking may disrupt semantic continuity and omit details within chunks, it offers a practical balance between preserving semantic integrity and enabling LLMs to comprehend specific details.

Sentence A sentence-level memory unit (Cheng et al., 2024) enables the capture of specific facts (Wang et al., 2023i) and details but tends to overlook connections between sentences, potentially leading to discontinuities in understanding. Thus, sentence-level segmentation is typically employed in scenarios requiring a detailed grasp of individual sentences' meanings, such as sentiment analysis (Guo et al., 2023b) and knowledge editing (Zhong et al.). In addition, MemPrompt (Madaan et al., 2022a) uses a sentence to assist the model in understanding specific tasks.

3.1.2 Graph

Graph-based data organization employs nodes and edges to structure knowledge systematically. Nodes represent distinct units of information, which can range from words and sentences to entire paragraphs. Edges, on the other hand, signify the relationships between these units, illustrating how they are interconnected. This structured approach to data organization proves particularly advantageous for tasks requiring advanced reasoning capabilities. For instance, reasoning through graphs enables flexible transitions between different reasoning paths, accommodating both logical deduction and multi-hop reasoning. Multi-hop reasoning involves traversing multiple interconnected nodes to infer complex relationships or derive conclusions that span diverse pieces of information. By leveraging this graph-based structure, one can achieve more efficient and nuanced reasoning processes. In a recent development, HippoRAG (Gutiérrez et al., 2025) incorporates the Personalized PageRank algorithm along with the inherent ability of an LLM to automatically build a knowledge graph, thereby enhancing the retrieval process with multi-hop reasoning capabilities.

Sub-graph Sub-graphs represent specific portions of the overall graph that are most closely related to the query (Ranade & Joshi, 2023). The nodes in a sub-graph may represent entities, sentences, or paragraphs. G-Retriever (He et al., 2024b) takes a textual graph describing nodes and edges as part of the input to use the structured information in the graph. HyKGE (Jiang et al., 2023) searches for sub-graphs based on the entities mentioned in the query and extracts the reasoning path to improve LLM's reasoning ability. The primary challenge of using sub-graphs lies in the high costs of constructing and maintaining sub-graphs.

Triple Triples are the fundamental units of graphs, representing entities and their relationships in a structured format (Wang et al., 2023f; Kang et al., 2023). They provide fine-grained knowledge but are limited by their fragmented nature, i.e., a triple only involves two out of a large number of entities, which may hinder the expression of continuous semantic information unless enough triples are retrieved. To use this precise memory, ISEEQ (Gaur et al., 2021)inserts triple facts into the appropriate position of the original query. KnowledGPT (Wang et al., 2023f) goes a step further and adds additional descriptions of the entities. Because this type of memory is concise and explicit, the retrieved irrelevant triples will directly affect the correctness of the generated content. Therefore, KALMV (Baek et al., 2023) proposes a detection method for this problem. Besides, PGMR (Sharma et al., 2025), a new modular architecture featuring a non-parametric memory retriever module for managing KG elements, thereby enhancing the accuracy and reliability of SPARQL queries generated by LLMs.

3.1.3 Vector

Vectors (Liu, 2022; Cheng et al., 2024; Chase, 2022; Yang et al., 2024a; Borgeaud et al., 2022; Do et al., 2024) characterize rich semantics and thus facilitate improved contextual understanding of LLMs.

Original text is segmented into smaller fragments that are then converted into vector representations through an encoding model and archived in a vector-based knowledge base. Typically given a query, recent works (Melz, 2023; Ren et al., 2023b; Maharana et al., 2024; Lin et al., 2023) retrieve relevant knowledge from the memory by computing the similarity between the query and the stored embeddings. The top K fragments with the highest similarity will be used as part of the prompt to improve the LLMs' understanding of the query.

As a memory format, vectors offer three **advantages** over free text and graph-based representations:

- **Robust semantic understanding** Vectors represent words, sentences, or even graph nodes in a continuous space where semantically similar entities are closer to each other. However, semantic relationships among free texts or entities in the graph are not apparently captured and need to be further inferred.
- **Scalability and Flexibility** Vectors can handle large-scale data with more efficient indexing, quantization, and batch processing, circumventing the limitations of simple keyword matching. The other two representations requires heavy pre-processing and more storage space suffering from higher complexity.
- **Generalization and Transfer Learning** Vectors capture underlying semantic similarity and can generalize to different tasks and domains with relatively little retraining. In addition, it is capable of processing multi-modal data, such as images and audio. Free-form text is typically task-specific, requiring manually crafted features, while graph structures are usually associated with specific topologies and require corresponding graphical algorithms.

3.2 Training with Explicit Memory

Training on explicit memory enables the model to efficiently utilize external memory by retrieving, adapting, and refining it, avoiding the need to reprocess the entire dataset. It involves learning how to interact, organize, and integrate well-structured memory representations during training rather than relying solely on retrieval based on memory similarity or structural matching, which can lead to redundancy.

There are several key **advantages** compared to training-free retrieval. Firstly, it optimizes retrieval relevance and generation accuracy by training each module for more context-sensitive use and alignment. Additionally, the training process enables LLMs to perform more sophisticated inferences by integrating external past experiences with the model's inherent reasoning capability. Specifically, for scenarios with long context that require consistency or accuracy across multiple turns of interactions, it ensures consistent responses over successive queries and adapt to rapidly evolving information or user-specific scenarios.

In this subsection, we distinguish two critical training phases: **pre-training** (§3.2.1) and **fine-tuning** (§3.2.2). We aim to explain how both stages contribute to the enhanced performance of models equipped with explicit memory systems. Besides, we introduce a few works to address the following three **main challenges**: i) **high computational costs** of real-time encoding of large-scale retrieved texts; ii) **lack of interpretability** in retrieval during training and inference; iii) **poor performance** in retrieval and generation when training autonomously. Tree typical pipelines for training with explicit memory can be seen in Fig. 7.

3.2.1 Pre-Training

Pre-training involves training an LLM from amounts of diverse text data by predicting the next token. The goal is to obtain an LLM with a comprehensive understanding of human languages and a general task-solving ability, thereby establishing a robust foundation for subsequent knowledge-intensive tasks and enhancing their retrieval capabilities.

Unsupervised Memory Retrieval REALM (Guu et al., 2020) proposed an approach to knowledge storage, successfully pre-training a knowledge retriever in an unsupervised manner for the first time. Specifically, it achieves end-to-end training by modeling both the retrieval and prediction processes. This work demonstrated significant performance improvements in open-domain question-answering tasks, introducing the first unsupervised pre-training method for a knowledge retriever using masked language modeling and backpropagation.

Advances in Retrieval Integration RETRO (Borgeaud et al., 2022), unlike REALM, directly appends retrieved content to the prompt and integrates retrieved data via chunked cross-attention. RETRO employs a pre-trained BERT as the retriever and freezes it during pre-training. This approach significantly enhances the memory capacity of LLM without increasing computational overhead, enabling the efficient integration of external knowledge at a much larger scale. Subsequently, RETRO++ (Wang et al., 2023a) analyzes RETRO and GPT models, illustrating the advantages of retrieval-augmented architectures in text generation and zero-shot knowledge-intensive tasks. Building on these insights, InstructRetro (Wang et al., 2024a) applied

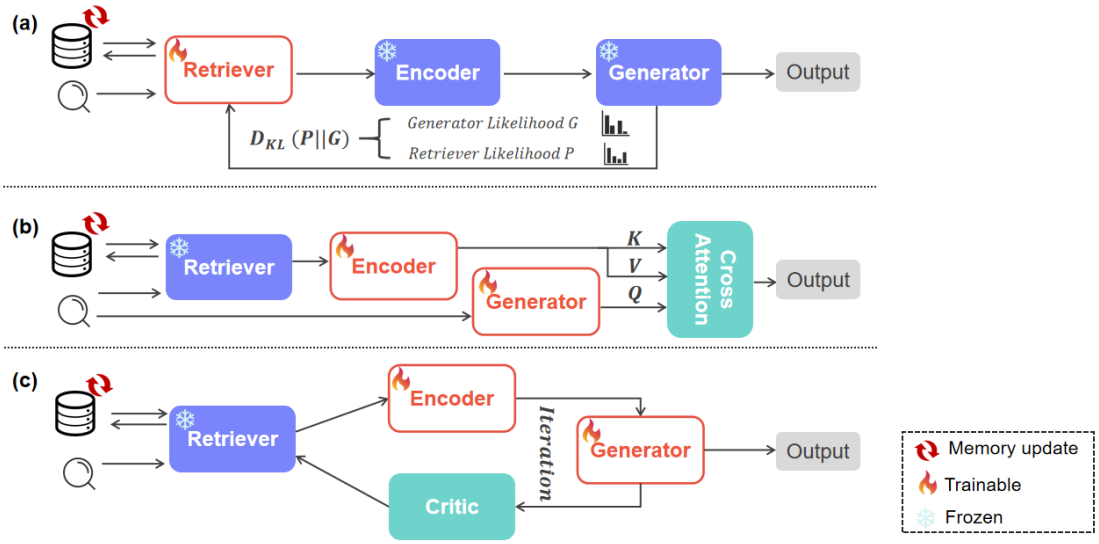

Figure 7: Three typical pipelines for training with explicit memory.

the RETRO framework to pre-training and instruction-tuning of GPT, resulting in improved accuracy and generalizability on complex, knowledge-intensive tasks.

Innovations in Knowledge-Intensive Pre-Training Memory^3 (Yang et al., 2024a) leverages a two-stage pre-training strategy by selectively storing key-value pairs with lower read and write costs. Memory^3 demonstrated its potential to enhance both the efficiency and performance of LLM across various tasks. Additionally, various pre-training methodologies have been proposed to address other challenges in LLM development. For example, Atlas (Izacard et al., 2023) employed a dual-encoder architecture for dense retrieval, combined with a sequence-to-sequence model, then used joint pre-training to better integrate retrieved documents for knowledge-intensive tasks. SANTA (Li et al., 2023b) utilized structured data alignment and masked entity prediction as pre-training techniques, achieving state-of-the-art performance in code and product search tasks. To maintain coherence and diversity in generation, CoG (Lan et al., 2023) generates text progressively copying fragments from existing corpora, outperforming traditional and retrieval-augmented models across multiple evaluations.

3.2.2 Fine-Tuning

Fine-tuning on the explicit memory allows the model to be specialized with task-specific or domain-specific knowledge, ensuring that it can handle nuanced information retrieval in targeted applications and scenarios.

A notable work is the Retrieval-Augmented Generation (RAG) model proposed by Lewis et al. (2020), which combines parametric memory from a pre-trained seq2seq model with non-parametric memory, such as a dense vector index of Wikipedia. By jointly finetuning the retriever and generator, RAG effectively learns to retrieve relevant information and condition its output on external knowledge, significantly improving factuality, diversity, and specificity in generated responses. This finetuning strategy has led to substantial performance improvements across knowledge-intensive tasks, such as open-domain question answering and fact verification.

Autonomous Memory Retrieval Building on RAG, Self-RAG (Asai et al., 2024) introduced a self-reflection mechanism that further enhances retrieval and generation processes. During training, a critic model generates reflection tokens, which are inserted into the training data. These tokens help the generator learn

when and how to retrieve relevant information more intelligently. By finetuning this self-reflective system, Self-RAG shows significant advantages in tasks that require factual verification, reasoning, and long-text generation, allowing the model to be more efficient and accurate in deciding when retrieval is necessary and how to use the retrieved content effectively. Besides, SMA (Zhang et al., 2025)introduces self-memory alignment to enhance the generalization of LLMs and balance trade-offs between different capabilities. Specifically, it fine-tunes the model on self-generated responses to precise, simple factual questions using preference optimization. Extensive experiments demonstrate that SMA significantly improves the overall performance of LLMs, consistently enhancing factual accuracy, helpfulness, and comprehensive skills across various benchmarks.

Context-Aware and Task-Driven Memory Retrieval To further improve task-specific performance, UPRISE (Cheng et al., 2023) finetunes a lightweight prompt retriever to automatically retrieve prompts tailored to specific inputs, improving LLMs' zero-shot capabilities of long-form question-answering. For context-aware dialogue generation, SURGE (Kang et al., 2023) leverages subgraph retrieval through graph-text contrastive learning. By finetuning both the subgraph retriever and the generator, SURGE enhances the model's ability to handle complex, structured data within dialogues.

Efficient Large-Scale Memory Retrieval As processing large amounts of retrieved text passages can increase computational costs, methods such as Fusion-in-Decoder (FiD) (Izacard & Grave, 2021) have been proposed to address these challenges. FiD processes each retrieved passage independently in the encoder while jointly processing them in the decoder, thus optimizing computational efficiency. Similarly, REPLUG (Shi et al., 2024) introduces a retrieval-augmented approach that integrates relevant documents with input context without modifying the internal parameters of the LLM. This reduces the computational burden of finetuning large models (e.g., 405B parameters) by minimizing the KL divergence between retrieval likelihood and the model's output perplexity.

3.3 Training with externalized parameteric knowledge

Externalized Parametric Knowledge effectively extracts and externally stores portions of the model's internal knowledge or its own intermediate outputs in a structured and accessible format.

It is particularly useful in tasks that involve processing **long documents**(§3.3.1) and **Knowledge injection**(§3.3.2), where the model's internal memory is insufficient for handling the entire context. It can retrieve the information as required, allowing for the efficient handling of extended contexts while maintaining coherence and accuracy throughout the task. This mechanism serves as a bridge between the model's ability to handle inherent knowledge and its ability to maintain and process large volumes of external information over time, avoiding the limitations of using internal memory alone. This method provides additional flexibility in accessing and utilizing information, ensuring that when the model's internal memory is not sufficient for a given task, explicit external storage comes to its rescue.

3.3.1 Long Contexts

The traditional Transformer architecture faces significant challenges in capturing long-range dependencies due to the limited context length imposed by the attention mechanism (Li et al., 2023a; Wu et al., 2025a). Yet, many tasks require models to process distant information, which is often crucial for accurate predictions. To address this limitation, Wu et al. (2022b) introduced MemTRM, a language model designed to memorize representations of previous inputs. MemTRM stores key-value pairs of past inputs and uses approximate k-nearest neighbor (kNN) search to extend the model's effective attention span. It demonstrates that MemTRM significantly improves performance across various tasks by expanding the model's attention context.

Building on this concept, several works take different approaches to extending the attention span. One typical framework leveraging external memory for long context can be seen in Fig. 8. One such model is Unlimiformer (Bertsch et al., 2023), which offloads cross-attention computation to a kNN index. Unlike MemTRM, Unlimiformer is fully non-parametric, requiring no fine-tuning, and allows each attention head

in every decoder layer to focus only on its top-k keys. This attention reconstruction capability enables Unlimiformer to perform personalized retrieval in each layer while maintaining greater efficiency than MemTRM. Another related work is FoT (Tworkowski et al., 2023), which extends the model's context length through fine-tuning rather than modifying the architecture. FoT has shown significant promise in enhancing the ability of LLM to handle long-text tasks effectively. A recent work EpMAN (Chaudhury et al., 2025) proposes an architecture combining episodic memory attention with self-attention during LLM training for robust long context performance.

Despite the advancements of MemTRM and its derivatives, the coupled memory design in MemTRM presents a challenge: the cached representations of past inputs may diverge from the current model representations as model parameters are updated. This distribution shift limits the effectiveness of memory-augmented models over time. To address this issue, LongMem (Wang et al., 2023e) decouples the network architecture by freezing the original LLM as a memory encoder and introducing an adaptive residual side network to act as the memory retriever and reader. This decoupling not only mitigates the issues caused by distribution shifts but also demonstrates superior performance in long-text processing and contextual learning tasks.

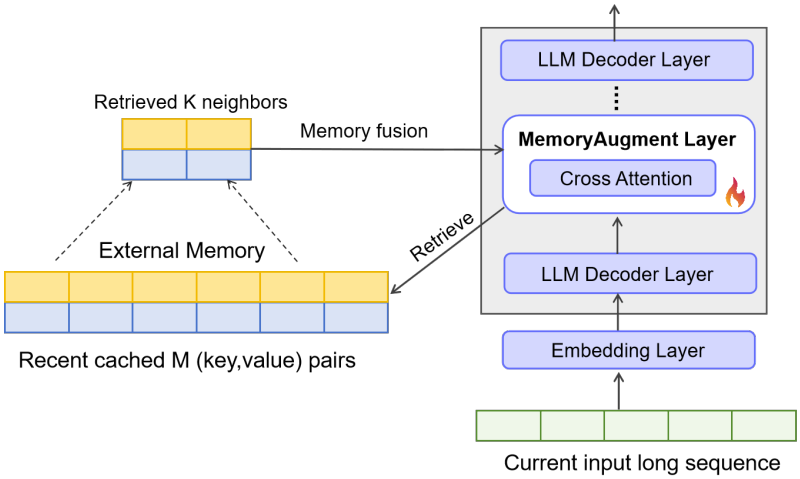

Figure 8: External memory training for long context.

3.3.2 Knowledge Injection

There are other research works leverages knowledge provided as part of the context from memory retrieval for knowledge injection and augmentation. These works aim to reason more robustly by folding the provided contextual knowledge into the model's parameters and complete the downstream tasks by using the updated parameters.

TRIME (Zhong et al., 2022a) utilizes a contrastive learning objective that aligns the hidden representation of a token with both token embeddings and a set of in-batch contextualized representations. The method introduces new strategies for memory construction and data batching to adapt to different memory types at testing time, which allows for back propagation to all memory representations. Similarly with various memory types, Yogatama et al. (2021) introduces an adaptive semi-parametric language model (SPALM) that integrates a non-parametric episodic memory component with extended short-term context through cached local hidden states and global long-term memory by retrieving nearest neighbor tokens at each timestep. A gating function is designed to adaptively combine information from local context, short-term memory, and long-term memory.

There are other works encoding text of different levels of granularity into embedding for retrieval. TOME (De Jong et al., 2021) integrates a semi-parametric representation into its architecture as a source of factual knowledge through "mention memory". It maintains a table of dense vector representations for every entity mentioned in a corpus, allowing TOME to retrieve and assimilate information from multiple sources making it more scalable and efficient compared. Open Predicate Query Language (OPQL) (Sun et al., 2021) uses a dual-encoder pre-training process to encode relation mentions, which can be integrated

into a language model (OPQL-LM) to improve performance on open-domain question answering. Unlike previous methods that rely on distant supervision from a knowledge base, OPQL is a method for constructing a virtual knowledge base from text without any structured supervision.

3.4 Limitations, open questions, discussion

To continuously learn, adapt, and improve the ability of LLMs to learn with explicit memory, we identify the following open questions and encourage future research to address them.

Can RAG solve the limitation of long context Enhancing the long context capabilities of LLMs is crucial in downstream tasks. There are currently two main approaches when handling longer sequences, extending an LMs context length or RAG. Training LLMs with extended context windows allows them to process long inputs continuously without truncation, thus improving their attention span. However, this also incurs higher computational costs and memory requirements. Methods using RAG dynamically retrieve concise, relevant documents based on the query, without the need to store the entire input. Further research is needed to develop mechanisms beyond simple semantic matching, which can create a more robust and dynamic memory system while avoiding memory overload. A deeper analysis of potential applications for each method, and how they can be integrated into a unified and flexible system, offers a promising direction for future research.

When and how to retrieve more intelligently and autonomously The development of more intelligent and autonomous memory retrieval mechanisms is crucial for optimizing tasks such as retrieval, reasoning and content generation. This enables the model to efficiently decide when retrieval is necessary and how to integrate the retrieved memory seamlessly for improved performance. An intelligent retrieval process is crucial for adaptively absorbing contextual information from external memory based on task-specific needs, especially for tasks requiring complex interactions with various data sources.

How to enhance retrieval-based training to avoid hallucination and contamination Incorporating explicit memory for knowledge injection during training allows LLMs to access relevant information more precisely and adaptively, rather than relying solely on parametric knowledge or context provided at generation time. However, LLMs may encounter memory contamination, where irrelevant or incorrect information is unintentionally stored during the learning process. A promising direction for future research is the design of selective memory mechanisms that filter out irrelevant details while retaining crucial information, thus reducing model hallucinations and preventing memory contamination.

In addition to these broad research topics, there are several engineering challenges that need to be addressed. For example, developing a retrieval method that maintains **consistency and coherence between external memory and the implicit memory** learned previously, without disruption over time. Incorporating large-scale retrieval-augmented models during training presents a significant challenge. The retriever must consider millions of candidate documents and backpropagate, which requires substantial computational resources. This presents an important direction for future work: optimizing the model structure, storage, and document selection to reduce computational burden and enhance **scalability and efficiency**.

4 Agentic Memory: Consolidating Memories into Humanic Agents

In the study of cognitive science, memory encompasses the cognitive processes involved in encoding, storing, and retrieving information. According to the Atkinson-Shiffrin three-stage memory model (Atkinson, 1968), information in the human brain progresses through three distinct stages: it initially enters sensory memory, then transitions to short-term memory, and ultimately consolidates into long-term memory, as illustrated in Figure 9.

LLM agents are artificial intelligence systems that understand and generate human-like text within real-world environments. These agents are capable of performing tasks such as answering questions and engaging in conversations. One of the central functionalities of LLM agents is their memory system, which mirrors the structure and processes of human memory. With memory modules, LLM agents can accumulate experiences, adapt continuously, and exhibit consistent, rational, and effective behaviors. Specifically, the memory modules

in LLM agents facilitate the retention of past interactions and knowledge, enabling the agents to reference previous information and improve their performance over time. In this context, we categorize the memory systems of LLM agents in a manner analogous to human memory, as follows:

Sensory Memory is the initial stage of memory, responsible for briefly retaining sensory information. It includes iconic (visual) memory, echoic (auditory) memory, and haptic (touch) memory. In the context of LLM agents, we regard Sensory Memory as the data ingestion pipeline of AI systems, such as DataLoader, and will not discuss it in detail.

Short-term Memory (STM) temporarily stores the information currently in awareness, as well as information necessary for complex cognitive tasks such as learning and reasoning[a]. For LLM agents, STM refers to the information maintained in the context window during in-context learning, which is thus constrained by the limited context window length of the Transformer architecture.

[a]Typically stores about 7 items, with a duration of approximately 20-30 seconds.

Long-term Memory (LTM) stores information for long periods in human cognition[a]. One type of LTM is declarative memory of facts and events, which can be consciously recalled. Another type is procedural memory including unconscious skills. For LLM agents, LTM serves as an external storage system, accessible to the agent during queries through efficient retrieval mechanisms.

[a]Typically with virtually unlimited capacity, lasting from days to decades

Although Zhang et al. (2024d) has examined the reasons, contents, and methods of storing memories in LLM agents, a comprehensive categorization of memory usage approaches based on an analogy to human cognitive processes, along with a review of engineering-level memory systems for agents available on the market, remains absent.

In this section, we will examine: (1) how an LLM agent's memory system operates across both **short-term and long-term** contexts (§ 4.1), drawing inspiration from human cognitive processes; (2) the mechanisms through which **multiple agents** share memories (§ 4.2); (3) the pipeline for **data ingestion, storing, indexing, and application** within agent memory systems (§ 4.3); and (4) methodologies for **evaluating** the effectiveness of these memory mechanisms (§ 4.4).

4.1 Single-agent Memory

Recently, several studies have been conducted to augment LLM agents with non-parametric external memory (Mai et al., 2023; Maharana et al., 2024; Yang et al., 2024a). This enhancement improves the agent's ability to explicitly store, retrieve, and utilize memory for various tasks. These works generally consist of two key stages: (1) recalling relevant thoughts from memory before generating a response, and (2) post-thinking after generating a response to incorporate both historical and new thoughts into memory, thereby improving consistency and efficiency.

In this subsection, we synthesize a wealth of research dedicated to leveraging external memory using refined and optimized RAG methods at inference time for the LLM agent on better downstream tasks.

4.1.1 Short-term Memory

Short-term memory serves as a transient storage system within LLMs, typically implemented by maintaining recent inputs within the context window. Due to the limited length of the context window, input history prior to a certain time point is inevitably discarded. Nevertheless, short-term memory enhances the continuity and consistency of LLMs and provides significant benefits in many challenging tasks, such as multi-hop reasoning and sequential decision-making.

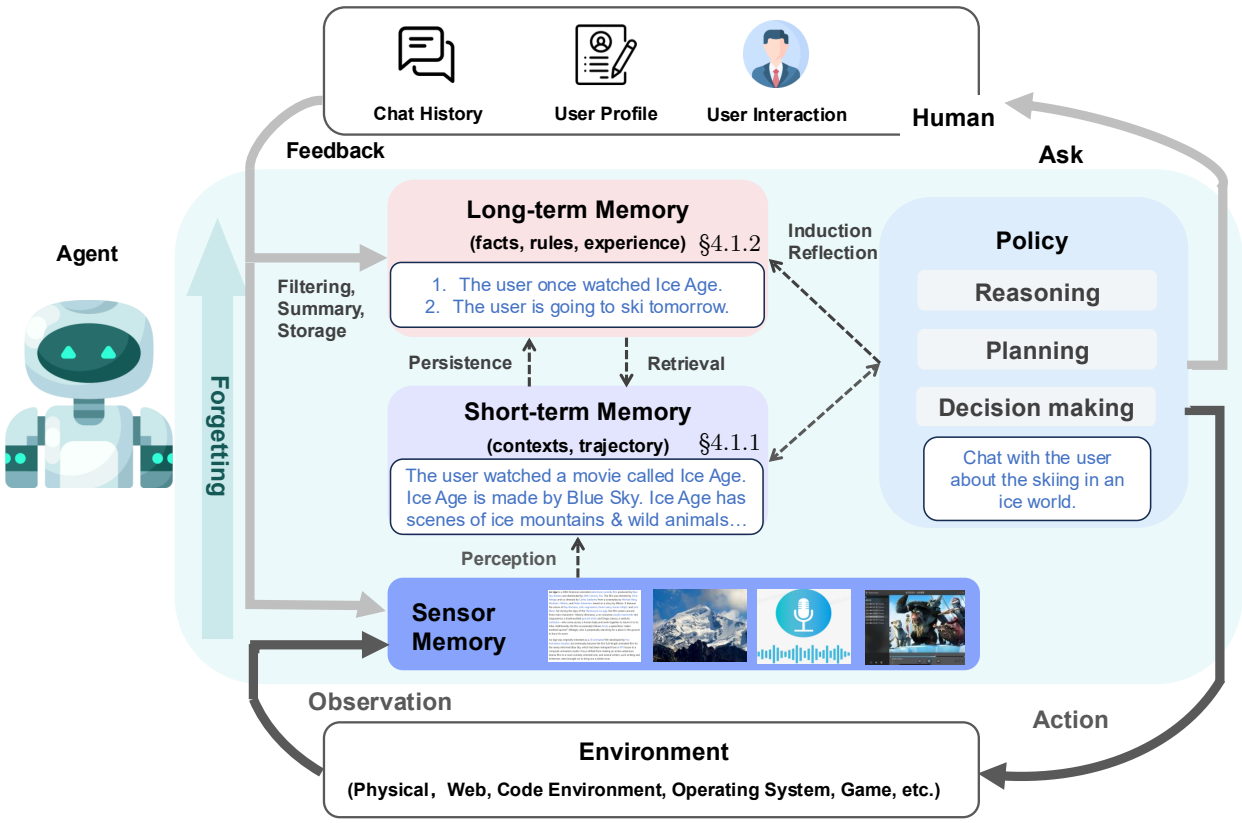

Figure 9: General memory architecture for a single agent. The agent interacts with the external environment (black arrows) and humans (gray arrows).

Chain-of-Thought (CoT) (Wei et al., 2022) prompting explicitly asks LLMs to generate intermediate reasoning steps that lead to the final answer. These intermediate steps (also known as thoughts) are stored in the context of LLMs, acting as their short-term memory and stimulating logical reasoning. **COT-SC** (Wang et al., 2022b) is an extension of the CoT framework that introduces a decoding strategy called self-consistency to enhance complex reasoning performance in language models by sampling diverse reasoning paths and selecting the most consistent answer. Unlike CoT, where reasoning follows a linear-chain structure, **Tree of Thoughts (ToT)** (Yao et al., 2024a) prompting solves problems by considering multiple reasoning paths organized in a tree structure. This tree-structured reasoning facilitates efficient forward-looking and backward-tracking, which are crucial elements of advanced search-based problem-solving. Building on CoT and ToT, **Graph of Thoughts (GoT)** (Besta et al., 2024) prompting enables reasoning over a graph structure, where each node represents an intermediate thought, and the edges encode the dependencies between them. GoT offers a more flexible prompting paradigm, as it supports a wide range of thinking transformations. For example, it allows convenient aggregation of thoughts and seamless switching between different reasoning flows within the graph structure.

ReAct (Yao et al., 2022) is a novel approach that prompts LLMs to generate both reasoning traces and task-specific actions in an interleaved manner, fostering synergy between the two processes. This method combines reasoning and acting of LLM and allows the model to dynamically update action plans through reasoning and interact with external sources like Wikipedia to enhance decision-making. **Reflexion** (Shinn et al., 2024) builds upon ReAct and enhances the decision-making ability of LLMs using verbal reinforcement learning. At the core of Reflexion is deliberate reflection on natural language feedbacks obtained from task outcomes, rather than through weight updates. Reflexion firstly explores the property of self-reflection in LLMs and shows that self-reflection is extremely useful to iteratively learn over trials as another form of short-term memory. Similar to Reflexion, which leverages linguistic feedback, an increasing number of works

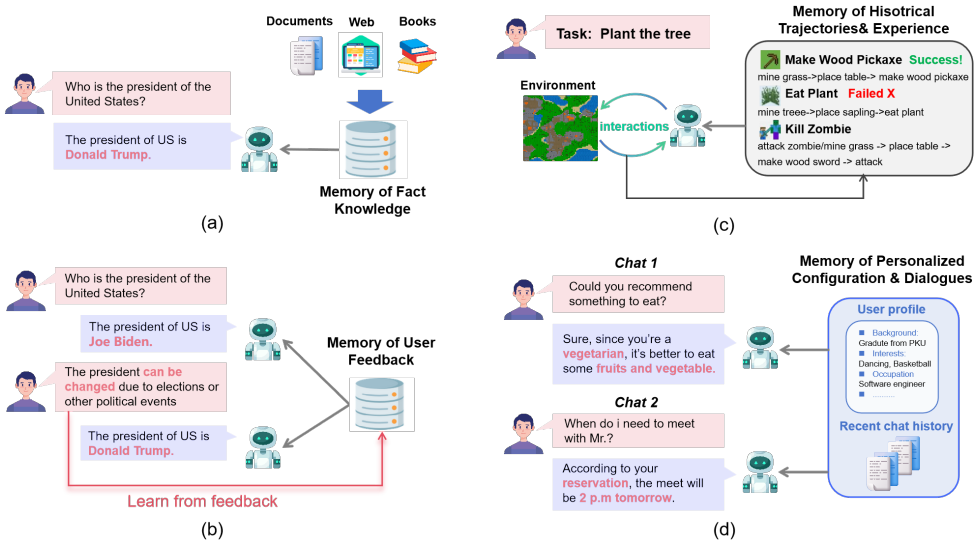

Figure 10: Long-term memory containing (a) fact knowledge; (b) historical trajectories and experience; (c) user feedback; (d) dialogues and personalized configuration.

have been proposed for self-improvement and evolution (Li et al., 2025b; Tang et al., 2024; Li et al., 2024b; Zhao et al., 2025). Gupta et al. (2024) learns general prompt instructions for LLMs using past self-reflections. Specifically, it gathers self-reflections in training and generalizes them into verbal 'meta-reflections' which serve as additional instructions to enhance the efficiency of agent's. RefAug (Zhang et al., 2024e) uses reflective augmentation to embed reflective sections within training instances, encouraging models to consider alternative solutions and engaging in deeper reasoning. Notably, This method goes beyond standard data augmentation by fostering a more comprehensive understanding of mathematical problems. Reflection on search Trees (RoT) (Hui et al., 2024) is introduced to reflect on an LLM's previous tree search experiences to generate guidelines, which are then used to improve the model's decisions in subsequent searches. This approach prevents repeated mistakes and enhances search efficiency. A key innovation is the identification of critical information from historical searches to produce more effective guidelines. Mirror (Yan et al., 2024) is a multiple-perspective self-reflection method to addresses the limitations of LLMs in self-assessment and feedback generation by introducing a Navigator-Reasoner framework. It improves through a heuristic interaction between a navigator, which provides question-adaptive directions, and a reasoner, which assesses and refines predictions based on the directions from the navigator. Textgrad (Yuksekgonul et al., 2024) is an automatic "differentiation" framework that optimizes composite AI systems by back-propagating textual feedback from large language models.

4.1.2 Long-term Memory

While operating in a real-time environment like Jia et al. (2024); Hafner (2021); Bellemare et al. (2013), crucial information, interactive feedback and distilled experience need to be continuously collected and further extracted from short-term memory and preserved as long-term memory of the LLM agent for future reference. To provide the agent with useful knowledge and experience for latter use, some works (Huang et al., 2024a; Park et al., 2023; Wang et al., 2023g; Zhang et al., 2024a) incorporate long-term memory (LTM) to LLMs from both user-specific and common-sense perspectives. This enables LLMs to flexibly utilize past experiences in accordance with current situations and enhance their task-planning and decision-making capabilities. These works usually formulate and organize thoughts in memory based on operations like insert, forget, merge and others, enabling dynamic updates and evolution of the memory in the long run.

As illustrated in Figure 10, this section focuses on four primary categories of long-term memory for an LLM agent:

Memory of fact knowledge The vast number of parameters in LLM endows them with remarkable capabilities, allowing them to excel in a variety of Natural Language Processing (NLP) tasks. However, this complexity also presents challenges, making LLMs difficult to train and inhibiting their ability to continuously assimilate new knowledge(Zhang et al., 2023g), which may lead to inaccuracies in their outputs. In order to resolve these issues, Du et al. (2023) propose a continual learning framework that incorporates memory mechanisms allowing LLMs to assimilate new knowledge and modular operators to enhance model inference with this newly acquired knowledge and dynamically adapt to evolving environments and continuously integrate new information without the need for parameter tuning. Similarly, Modarressi et al. (2023) augments LLM with a write-read memory module by extracting and saving knowledge in the form of triplets, allowing for scalable, updatable, interpretable, and aggregatable memory storage, which is particularly beneficial for handling temporal-based question answering tasks.

Memory of historical trajectories and experience Incorporating memory into agents is crucial for enhancing their ability to remember historical trajectories, which ultimately improves decision-making and learning efficiency. By allowing agents to recall past experiences, they can identify patterns, make more informed predictions, and adapt their strategies based on historical data. This memory mechanism enables agents to build upon previous knowledge, facilitating more nuanced and contextually relevant responses. Delving into the integration of memory also supports the development of more sophisticated models that parallel the cognitive processes of humans, thereby advancing the overall effectiveness and versatility of artificial intelligence systems. Guo et al. (2023a) incorporate a centralized working memory hub and episodic buffer access to retain memories across episodes. This architecture aims to provide greater continuity for nuanced contextual reasoning in intricate tasks and collaborative scenarios. Liu et al. (2023) also maintains an evolved memory for storing historical thoughts and employs Locality-Sensitive Hashing for efficient retrieval. Kagaya et al. (2024) introduces Retrieval-Augmented Planning (RAP) with a contextual memory module to leverage past experiences for improved decision-making in complex tasks. RAP dynamically retrieves relevant past experiences based on current context to guide planning and action selection, mirroring human-like analogical reasoning.

Memory of user feedback Recently, many researchers have proposed various methods for LLM to continue to improve without retraining. The Language Model (LM) is coupled with a growing memory module with feedback either corrections on the historical errors or better clarifications on the tasks from the users. Madaan et al. (2022b) maintains a growing memory of misinterpretation of user intents, along with user feedback for clarification. This memory enables the system to produce enhanced prompts for new queries based on past user feedback, effectively leveraging previous corrections to improve performance on similar tasks. Tandon et al. (2021) enables LLM's to improve their output after deployment without retraining, by leveraging user feedback. It maintains a dynamic memory of cases where users have identified and corrected output errors, and uses a trained corrector model to apply similar feedback to fix new errors. This approach shows significant improvement in repairing errors and avoiding past mistakes on new examples. The system represents a step towards continuous model enhancement through interactive learning and memory-based feedback reuse. Dalvi et al. (2022) integrates a dynamic memory component that stores user-provided corrections to the model's erroneous beliefs. These corrections are retrieved and used as additional context when answering new questions, helping the system avoid repeating past mistakes. This approach represents a novel application of memory-based continual learning for belief maintenance in language models, allowing for user-driven system enhancement over time. In order to maintain an ever-improving memory for LLM, Li et al. (2024b) utilizes recursive reasoning-based retrieval and experience reflections to continually update the memory and learn from communicative feedback provided by users. Without periodically re-training, it enables LLM to obtain fresh knowledge and historical experience by dynamically improving and growing a continually updated memory through human communications.

Memory of dialogues and personalized configuration To address the limitation of context capacity over long conversations, Aadhithya A et al. (2024) introduces a novel memory structure that recursively aggregates dialogue context flexibly to enhance long-term memory for dialogue agents. It allows for broad coverage of information with controlled depth through conditional tree traversals, balancing the breadth and depth of information for long-form dialogues, which is crucial for multi-turn reasoning without exponential

parameter growth. Based on this, Chen et al. (2024a) adopts compressive memory that integrates session-specific summaries, user-bot dynamics, and past events into a concise memory format. This method is designed to be more manageable and efficient than traditional retrieval-based methods using Direct Preference Optimization (DPO) to enhance the model's ability to generate contextually appropriate responses. In Maharana et al. (2024), memory in these papers contains more nuanced and human-like conversational experiences, showing its effectiveness in managing time-dependent information and maintaining coherence in long-term and varied interactions, significantly contributing to the field of conversational agent.

Incorporating personalized knowledge bases into memory allows users to effectively store and access specific knowledge according to their requirements. Wang et al. (2023f) is a novel framework for knowledge retrieval and personalized knowledge base interaction. It employs the "Program of Thoughts" (PoT) prompting method, which facilitates model interaction with Knowledge Bases through the generation of Python code, thereby enabling knowledge retrieval. For domain specific tasks, Zhang et al. (2023c) introduces a personalized medical assistant tasks through a computational bionic memory. It utilizes a memory generation module using Dual-Process enhanced Memory (DPeM) that fine-tunes the LLM to produce personalized responses. Zhong et al. (2024) enables storing past conversations and adapting to user personalities. The system is showcased through SiliconFriend, a chatbot that provides empathetic and long-term companionship, demonstrating MemoryBank's effectiveness in improving AI engagement.

4.2 Multi-agent Memory

In LLM-based multi-agent collaboration, shared memory mechanisms enhance agents' ability to leverage historical information, improving reasoning and coordination. Prior works vary in focus, with some emphasizing real-time memory sharing for efficient reuse and others prioritizing agent autonomy by exchanging only essential information for flexibility and robustness, posing the challenge of balancing experience collection and information overload.

Conceptually, shared memory serves as a repository of past interactions and knowledge that agents query to inform current actions, often implemented as vector-based representations with similarity metrics retrieving relevant memories. Recent frameworks like A-MEM (Xu et al., 2025), DAMCS (Yang et al., 2025), MS (Gao & Zhang, 2024), and IoA (Chen et al., 2024b) offer innovative approaches to shared memory in multi-agent systems. Drawing inspiration from the Zettelkasten method, A-MEM emphasizes memory organization and evolution, creating interconnected knowledge networks through dynamic indexing and linking, where comprehensive notes with structured attributes evolve over time to refine contextual understanding and improve performance in long-term tasks and multi-hop reasoning. DAMCS enables decentralized cooperation through hierarchical knowledge graphs, dynamic team formation, and structured communication, allowing agents to learn from each other's experiences and adapt to new environments via most relevant information, while keep their own individual memories. MS focuses on real-time memory sharing by storing Prompt-Answer (PA) pairs in a shared memory pool, with an autonomous retriever ensuring relevant memories enhance response quality and reduce dependence on external databases. IoA adopts an Internet-inspired framework, enabling seamless integration of heterogeneous agents via an instant-messaging-like architecture, facilitating agent discovery, dynamic team formation, and structured communication for scalable collaboration.

4.3 System Architecture

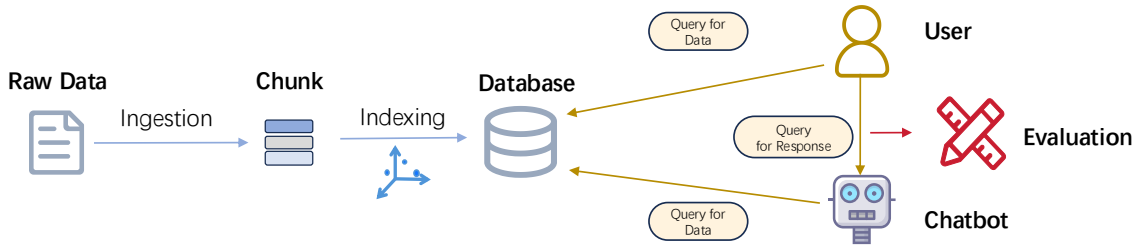

Figure 11: A general architecture of memory-augmented agent pipelines.

To enable LLM with external memory capabilities, a variety of open-source tools and frameworks have been developed. As shown in Figure 11, these systems typically consist of the following modules: **Data Ingestion → Storage and Retrieval → User Interfaces and Application Invocation**. In this section, we compare the existing mainstream agent system architectures module by module. A summary of these comparisons is shown in Table 2. Notably, evaluation of the memory systems is also an important module, which is discussed in the next subsection.

Table 2: Comparison of Tools for Data Ingestion, Storage, and User Interfaces

Tool	Data Ingestion	Storage and Retrieval	User Interfaces
Common	• Universal connectors (files/web/DBs)	• Vector DB • Hybrid indexing	• API access • Logging tools
MemGPT	• Conversation context • Tool execution outputs	• Hierarchical tiers	• Notebook examples
Zep	• Temporal streams • Agent message graphs	• Graphiti engine • Relationship expiration	• Dashboard monitoring • Graph visualization
Dify	• Web-based upload • Bulk knowledge import	• External vector DB • Response annotations	• No-code builder • Memory edit UI
Mem0	• Cloud sync • Cross-session data	• Managed backend • Versioned storage	• Personalized UI • Memory triggers
Haystack	• SQL/API ingestion • Streaming pipelines	• Hybrid search	• REST API • Pipeline configurator
LangChain	• Modular loaders • Notion/Slack integration	• Conversation buffer	• Memory inspection • Chain debugging
LlamaIndex	• 160+ formats • Structured data parsing	• Tree-Index	• Query analyzers • Index visualization

4.3.1 Data Ingestion

Comprehensive frameworks such as LangChain[3], LlamaIndex[4], and Haystack[5] provide extensive data connectors that support a wide range of data sources. These include local files (e.g., TXT, PDF, Word), web scraping, database integration, APIs, and streaming data. For example, LlamaIndex natively supports over 160 data formats, covering plain text, tables, SQL databases, notion documents, Google Drive, Slack messages, and API responses, making it highly versatile for various business scenarios. These frameworks often allow customization of data preprocessing, such as specifying chunk size, filtering special symbols, and extracting metadata to optimize indexing efficiency.

Conversational and agent memory focus mainly on data generated from interactions. For instance, **Letta (MemGPT)**[6] emphasizes conversation context and tool usage results, with data sources including user messages, model responses, and outputs from external tools (e.g., code execution results). Text retrieved from a search tool may be incorporated as new memory. In multi-agent scenarios, data sources additionally include messages exchanged between agents.

Domain-specific knowledge-based applications, such as customer service and knowledge retrieval systems, rely on data sources such as product documentation, procedural manuals and FAQs. Tools like Haystack facilitate efficient document import, enabling bulk loading and automatic conversion into internal indexing formats.

[3]https://www.langchain.com/
[4]https://www.llamaindex.ai/
[5]https://haystack.deepset.ai/
[6]https://www.letta.com/

4.3.2 Storage and Retrieval

Vector databases are the most prevalent approach, which emphasizes large-scale semantic query performance. Many frameworks integrate vector libraries such as FAISS[7], Chroma[8], Pinecone[9], Weaviate[10], and Milvus[11] by default. LangChain and LlamaIndex use local vector storage (either in memory or on disk) to store embeddings and can optionally connect to external service-based vector databases for persistence and distributed capabilities. In vector-based storage, text is segmented and encoded in high-dimensional vectors using models like Sentence-BERT (Reimers & Gurevych, 2019). The index consists of these vectors, which are matched with query vectors during retrieval to return relevant results. Retrieval hyper-parameters, such as top-K value or similarity threshold, can be fine-tuned. Some tools employ additional strategies, such as summary or hierarchical indexes. LlamaIndex's Tree-Index organizes documents into a hierarchical structure, enabling efficient retrieval for long-document QA. Keyword or hash indexing can also complement vector searches to improve accuracy and speed.

Graph databases excel in structured knowledge representation, logical reasoning, relationship tracing, and version control. Systems like Zep[12] parse conversations into nodes and edges with attributes (e.g., context summaries, embeddings, timestamps). Its Graphiti engine dynamically updates the graph, invalidating outdated relationships. Retrieval utilizes graph algorithms and explores node relationships. Mainstream frameworks like LlamaIndex and LangChain also support graph databases for enhanced knowledge management.

Hybrid storage strategies are sometimes employed to meet diverse needs, balancing speed and capacity. For instance, Letta introduces hierarchical storage, keeping recent conversations in immediate context while compressing older information into external databases for long-term archiving. Persistence is another major consideration. If vector indices stored in memory are not saved, they are lost upon restart. Hence, most frameworks offer options for persisting indexes and memory, such as dumping indexes to disk files or directly to clouds. Mem0[13] provides cloud-hosted services, allowing developers to store memory in its managed backend for durability and cross-session sharing.

Traditional databases and file storage are also used in simpler cases, which may rely on key-value stores (e.g., Redis), relational databases, or local files to record full conversations. However, such approaches struggle to scale to complex scenarios and are gradually being replaced by vector databases.

Information retrieval methods vary by application needs: chatbots prioritize recent conversation context and related knowledge, often using sliding windows (e.g., LangChain's ConversationBufferWindowMemory), while knowledge QA systems focus on pinpointing accurate answers from large databases, favoring semantic search combined with cross-verification. The organization of retrieval outputs is crucial; frameworks often truncate or filter retrieved memories to fit within context windows, with methods allowing LLMs to generate summaries from multiple sources. Graph community summaries similarly aggregate node content into concise contexts, ensuring prompts contain essential information without exceeding length limits.

4.3.3 User Interfaces and Application Invocation

Developer frameworks like LangChain, LlamaIndex, and Haystack provide APIs and libraries for integration but lack graphical user interfaces (GUIs). Memory functions operate in the background, allowing developers to review retrieval results through logs or debugging tools. API-based services like Zep store and retrieve conversation history through API calls, operating invisibly to end-users. Developers monitor usage via dashboards that track stored conversations and vector indexes. Command-line tools and developer utilities are available in some open-source projects. For example, MemGPT provides Jupyter Notebook examples for API-based memory retrieval.

[7] https://faiss.ai/
[8] https://www.trychroma.com/
[9] https://www.pinecone.io/
[10] https://weaviate.io/
[11] https://milvus.io/
[12] https://www.getzep.com/
[13] https://mem0.ai/

Full-featured platforms such as Dify[14] and Mem0 offer web-based UIs for chatbot configuration, knowledge base management, and real-time interactions. Dify enables non-programmers to build memory-enabled bots visually, while Mem0 provides a hosted ChatGPT with persistent memory, allowing users to upload knowledge and personalize interactions.

4.4 Evaluation on Agent Memory

4.4.1 Characteristics of memory

Understanding the characteristics of the agent memory is crucial for designing systems that effectively store, retrieve, and utilize information. In this subsection, we highlight the role of memory in optimizing performance, ensuring efficiency, and enabling continuous learning for an LLM agent. By examining these features, we aim to provide a comprehensive overview of memory's multifaceted nature and its impact on system behavior, as shown in Table 3:

Table 3: Characteristics of memory.

Feature	Type	STM	LTM	Subjective eval	Objective eval	Description
Temporality	Direct	✓	✓	✓	✗	Whether the memory includes time mentions, timestamp, temporal-based long dependency correlations
Consistency	Direct	✓	✓	✓	✗	Whether the memory remains to be consistent or not
Redundancy	Direct	✓	✓	✓	✗	Whether the memory maintains redundant information or not
Variance	Direct	✗	✓	✓	✗	Whether the memory is static or dynamic that can be updated
Transformation	Direct	✗	✓	✓	✗	Whether the memory can be converted between short-term and long-term

Temporality indicates whether temporal facts or events exist in the memory and how they are stored, accessed, and utilized over time. Memory with time-sensitive information is crucial for tasks involving temporal understanding, reasoning and long dependency tracking, which enables the model to distinguish information in a timeline order. **Consistency** reflects whether the memory remains stable and consistent across interactions. It ensures the credibility and reliability of the LLM agent when generating the output based on relevant memory in context. In the meantime, it avoid frustration when retrieving conflicting information across different queries. **Redundancy** indicates whether the memory maintains redundant information, such as storing multiple versions of the same fact or event. While redundancy can serve as a backup for fault tolerance, excessive redundancy can lead to inefficiency and confusion. **Variance** refers to the memory to be dynamic that it can be updated and merged with new incoming information without overwriting critical old data or losing coherence. Dynamic updating is critical for LLM agents in real-time applications, such as interactive agents or systems that must adapt to new facts or corrections. **Conversion & Transformation** refers to whether memory can be transferred between short-term memory (STM) and long-term memory (LTM) effectively. Real-time interactions benefit from STM, whereas LTM supports knowledge retrieval for tasks requiring continuity over sessions, comprehensive reasoning and continuous learning.

4.4.2 Capabilities of memory

In this section, we propose two types of test type to evaluate the capability of an agent memory as below:

- **Functionality Test (FT)** is a type of black-box testing that evaluates whether the software system performs its intended functions correctly as expected according to the defined requirements. It focuses

[14]https://dify.ai/

on verifying the functional correctness of the application by checking specific features and operations about the memories. Here we use FT to test capabilities like *Learning Efficiency* and *Generalization*.

- **Perturbation Test (PBT)** evaluates the stability of a software system by introducing modifications or disturbances to its inputs, environment, or internal state. The goal is to discern how the system reacts to perturbations of the memories and whether it can maintain the expected behavior. Here we use PBT to test capabilities like *Controllability* and *Robustness*.

The detailed definitions of the evaluation metrics mentioned above can be found in Table 4. We conducted these tests based on RetrievalQA Zhang et al. (2024f) dataset using four different types of agents, all using Llama3-8B-IT(Grattafiori et al., 2024) as the base model, with varying memory settings:

- In-Context Learning (ICL): Memories are entirely stored in the LLM's context.
- Retrieval-Augmented Generation (RAG): Standard RAG for storing memories.
- RAM: Continuously improving memory based on external feedback, implemented according to Li et al. (2024b).
- General Agent: An agent that automatically decides how to handle external feedback at each step, implemented using the Concordia framework (Vezhnevets et al., 2023).

Table 4: Memory capabilities metrics.

Test type	Capability	Type	STM	LTM	Sub. eval	Obj. eval	Description
FT	Learning Efficiency	Indirect	✗	✓	✗	✓	Task performance increase with accumulative memory
	Generalization	Indirect	✗	✓	✗	✓	Unseen task performance with the memory learnt from historical tasks
PBT	Controllability	Indirect	✓	✓	✗	✓	Task performance provided with unknown or counterfactual contexts
	Robustness	Indirect	✓	✓	✗	✓	Task performance provided with irrelevant contexts as noise

In Figure 12, we illustrate the performance of each memory setting across varying context quantities, ranging from 10% to 100% of the memories required to answer the question. This analysis highlights the learning efficiency of RAG and RAM, as performance basically improves consistently as the proportion of necessary memories increases.

Table 5 summarizes the evaluation results for Generalization, Controllability, and Robustness. The applied memory perturbations significantly influence performance, demonstrating the sensitivity of these metrics to changes in memory inputs. However, the generalization ability of past memories has not been observed, potentially due to insufficient data volume.

Table 5: Evaluations on capabilities of memory using different agents. M^{upd} indicates that the ground truth of the current question will be updated into the memory.

	ICL		**RAG**		**RAM**		**Concordia Agent**	
	wo M^{upd}	**w** M^{upd}	**wo** M^{upd}	**w** M^{upd}	**wo** M^{upd}	**w** M^{upd}	**wo** M^{upd}	**w** M^{upd}
Generalization	20%	20%	38%	38%	52%	52%	16%	16%
Controllability	20%	16%	54%	48%	38%	44%	16%	24%
Robustness	20%	4%	38%	19%	54%	26%	16%	12%

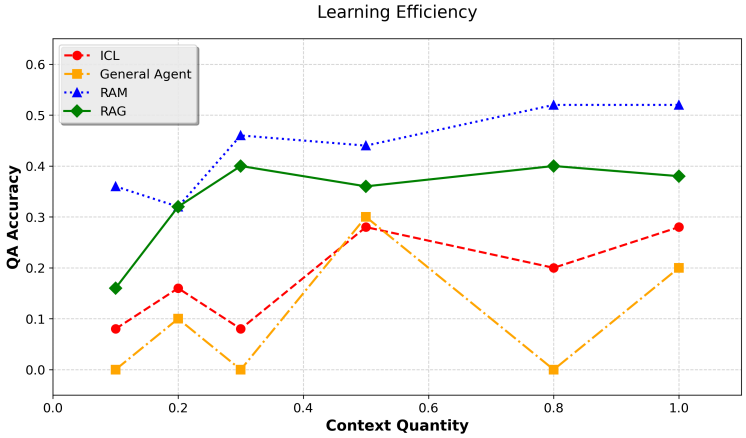

Figure 12: Learning Efficiency: Question-answering accuracy as a function of the proportion of necessary context provided to the LLMs.

4.5 Limitations and Future Works

To enhance the utility of agents, enabling dynamic memory adaptation during reasoning and communicative learning could be crucial. Inspired by human pedagogy, methods like RAM (Li et al., 2024b) demonstrate the potential of recursive retrieval and experience reflection for continuous memory updates based on user feedback. Additionally, in multi-agent systems, ensuring adaptive network structures and robust communication frameworks (Mao et al., 2024; Marro et al., 2024; Liu et al., 2024b) to facilitate effective memory synchronization also remains as a challenge.

5 Memory-augmented Multi-Modal Large Language models

Addressing the complexities inherent in multimodal context modeling, a critical research inquiry emerges:

How can we devise and execute memory mechanisms that adeptly amalgamate and preserve extensive multi-modal contextual data, thereby augmenting the comprehension and manipulation of intricate datasets within fluid environments?

This question is particularly salient within the domains of vision and robotics, where the optimization of contextual memory is paramount for bolstering the cognitive and functional capacities of embodied agents. Such advancements would empower these agents to execute sophisticated operations, enabling systems to make informed decisions based on comprehensive contextual understanding.

5.1 Multimodal Context Modeling with Memory

In this section, we introduce the multimodal context modeling with memory, incorporating information from audio, video, and other modalities. For each modality, we will discuss its modeling in relation to various downstream tasks.

5.1.1 Audio Context Modeling

The continuous and high-frequency nature of audio presents significant challenges in efficiently modeling its sequences, demanding substantial computational resources. As a result, developing effective methods to incorporate audio history is crucial for various applications. Recent advancements in audio context modeling have introduced innovative solutions to address these challenges. For instance, Conformer-NTM (Carvalho & Abad, 2023) proposes an external memory network between the encoder and decoder transformers for automatic speech recognition (ASR), enhancing the system's ability to handle complex audio sequences. Similarly, Loop-Copilot (Zhang et al., 2023d) introduces a Global Attribute Table that identifies and manages

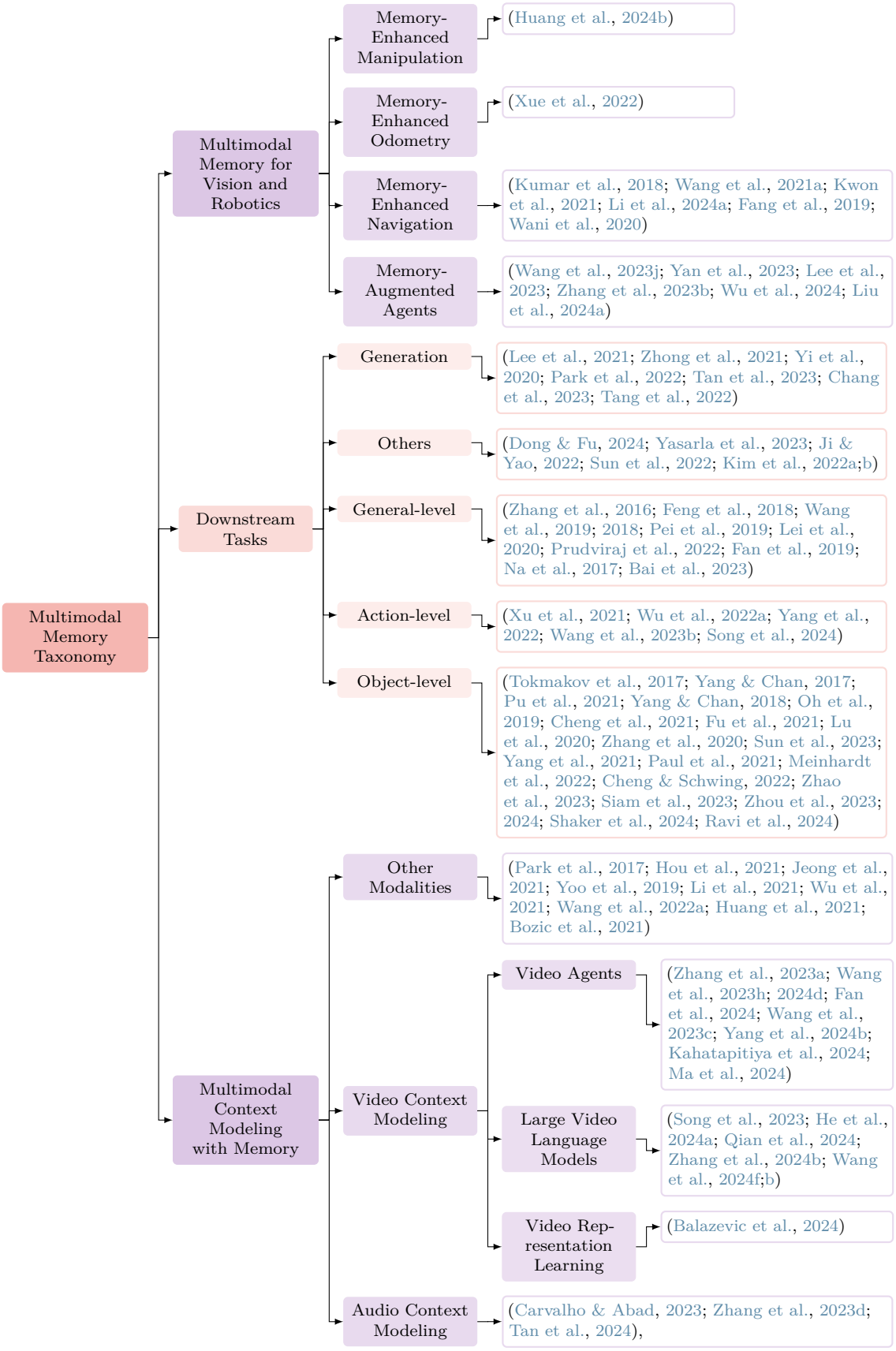

Figure 13: Taxonomy of multimodal memory applications and downstream tasks.

various musical attributes at any moment, aiding in task execution and ensuring musical coherence in music generation. Furthermore, MR-MT3 (Tan et al., 2024) employs previous instrumental tokens as key/value memory, effectively preventing instrument leakage in automatic music transcription tasks. These approaches demonstrate promising strategies for improving audio modeling, paving the way for more efficient and coherent audio processing systems.

5.1.2 Video Context Modeling

Video context modeling presents unique challenges compared to image processing due to its added time dimension, which significantly increases complexity. Most existing approaches use sampling-based methods, converting continuous video into discrete stacked frames. This makes balancing computational complexity with detailed video information a crucial focus in video research. Memory mechanisms have emerged as a vital strategy in achieving this balance. This section explores various aspects of video context modeling, starting with general video representation learning, focusing on developing more effective video encoders. We then delve into recent advancements in memory mechanisms for large video language models and video agents, which demonstrate strong performance in zero-shot video-language benchmarks and applications. Finally, we discuss the role of memory in various downstream tasks.

Memory-enhanced Video Representation Learning In the realm of video representation, MC-ViT (Balazevic et al., 2024) introduces a long video encoder using memory consolidation and cross-attention on video segments to efficiently encode long-context videos. This enhances the ability to handle extensive video data while maintaining detailed representation.

Large Memory-enhanced Video Language Models For large video language models, several innovative memory-augmented approaches have been developed for long video modeling. MovieChat (Song et al., 2023) builds on Q-Former for visual feature extraction with memory consolidation to model long videos for video question answering (QA). MA-LMM (He et al., 2024a) extends this with a retrieval strategy based on the semantic similarity of frame features. VideoStreaming (Qian et al., 2024) proposes an method combined with an adaptive memory selection strategy on the recurrent image feature, selecting a constant number of question-related memories using Gumbel-Softmax. Flash-VStream (Zhang et al., 2024b) presents a hierarchical memory system, incorporating FIFO queue for spatial features, and uses abstract memory implemented by cross-attention for whole video modeling. VideoLLaMB(Wang et al., 2024f) introduces a recurrent memory bridge with a memory cache to model video history in memory for long video understanding. Additionally, OmniDrive (Wang et al., 2024b) utilizes a memory bank for frames in autonomous driving-related QA.

Memory-enhanced Video Agent Video agents have made significant strides by transforming various video elements into text, such as captions, object names, and timestamps, which are then stored as external memory for LLMs to enhance video processing tasks. LLoVi (Zhang et al., 2023a), LifelongMemory (Wang et al., 2023h) and VideoAgent(Wang et al., 2024d) leverage captions as a memory for LLMs. Furthermore, VideoAgent (Fan et al., 2024) combine captioning, tracking, VQA modules to inject video object detection ability into LLM. ChatVideo (Wang et al., 2023c) utilizes captioning, tracking, and audio modules to extract information from video as memory. DoraemonGPT (Yang et al., 2024b) adopts a more comprehensive method by incorporating captioning, tracking, ASR, detection, action, and segmentation module to provide additional information for LLMs in video QA and segmentation tasks. LangRepo (Kahatapitiya et al., 2024) integrates captions, and timestamps as memory to LLMs for sequential understanding. Finally, DrVideo (Ma et al., 2024) reinterprets long-video understanding as a long-document comprehension task, effectively utilizing the power of large language models. These advancements underscore the diverse strategies employed to address the complexities of video context modeling and suggest promising directions for future research.

5.1.3 Other Modalities

Beyond audio and video context modeling, memory strategies are increasingly leveraged across various modalities to enhance context understanding and improve performance. In image captioning, the CSMN model (Park et al., 2017) enhances memory networks by using them as repositories for multiple types of context information, appending previously generated words to capture long-term information, and employing

a CNN memory structure for better context representation. For anomaly detection, the DAAC model (Hou et al., 2021) modulates reconstruction capabilities by generalizing the memory module in a blockwise manner using a multi-scale approach. In image-to-image translation, MGUIT (Jeong et al., 2021) explores memory networks to improve translation results. MemoPainter (Yoo et al., 2019) introduces a memory-augmented colorization model that achieves high-quality colorization with limited data. For blind face restoration, RMM (Li et al., 2021) proposes a wavelet memory module that stores spatial features of low-quality images and guides high-quality restoration. In semantic segmentation, a memory-based approach (Jin et al., 2021) stores significant training image representations, while MM-Net (Wu et al., 2021) uses learnable memory embeddings for few-shot segmentation, and CDFSS (Wang et al., 2022a) employs a meta-memory bank to bridge domain gaps. In deraining, MOSS (Huang et al., 2021) uses a self-supervised memory module to record prototypical rain patterns. Lastly, for 3D scene reconstruction, TransformerFusion (Bozic et al., 2021) proposes a hierarchical memory of input frame features for online video 3D scene reconstruction. These applications underscore the versatility and effectiveness of memory networks in improving context modeling across diverse tasks.

5.2 Downstream tasks

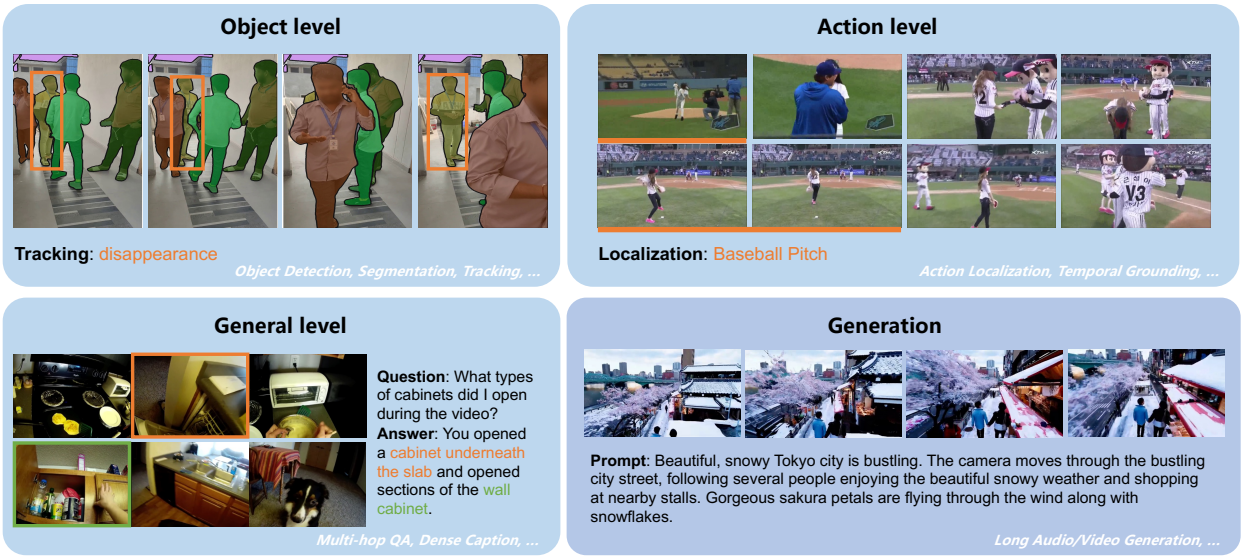

Figure 14: Various video understanding and processing tasks.

Advancements in video understanding and generation have highlighted the critical role of memory mechanisms in enhancing long-context modeling capabilities. Researchers have been increasingly focused on developing innovative approaches to leverage these memory mechanisms across various video processing tasks. This survey provides an overview of these advancements, beginning with video understanding at the object level, progressing through action-level tasks, and culminating in general-level applications such as summarization, captioning, and question answering. Figure 14 provides some tasks demonstration. Additionally, we explore recent breakthroughs in video generation that utilize memory-based architectures.

Object-level Task Object segmentation and tracking in videos are foundational tasks that benefit significantly from memory-augmented models. Early efforts in this domain employed RNNs, LSTMs, and GRUs to maintain and update memory states over time. For instance, ConvGRU (Tokmakov et al., 2017) integrates convolutional gated recurrent units to track the evolution of objects within a scene, while RFL (Yang & Chan, 2017) utilizes convolutional LSTMs for object tracking with preserving video instory. RMAN (Pu et al., 2021) builds on the LSTM architecture by adding a memory activation layer specifically for visual tracking.

Further advancements introduced more sophisticated memory networks. MemTrack (Yang & Chan, 2018) employs a dynamic memory network to store and recall target information, utilizing an LSTM to manage

memory operations for template-matching tasks. STM (Oh et al., 2019) and its enhanced version, STCN (Cheng et al., 2021), compute spatio-temporal attention across video frames, improving pixel-level object distinction. STMTrack (Fu et al., 2021) introduces a mechanism that stores historical target information to guide the tracker toward the most informative regions.

Graph-based memory networks have also shown promise in video object segmentation. GraphMemVOS (Lu et al., 2020) leverages an episodic memory network structured as a fully connected graph, facilitating cross-frame correlation capture. DTMNet (Zhang et al., 2020) builds on this by incorporating both short- and long-term memory storage to enhance temporal modeling. TMRN (Sun et al., 2023) improves memory retrieval operations by spatially aligning memory frames with current frames before temporal aggregation.

The advent of transformer-based architectures has further revolutionized object segmentation and tracking. AOT (Yang et al., 2021) employs long-short term memory for associating multiple object segments, while IMANet (Paul et al., 2021) utilizes attention mechanisms to access semantic information stored in memory. TrackFormer (Meinhardt et al., 2022) and XMem (Cheng & Schwing, 2022) exemplify the integration of transformers with memory modules to handle occlusions and segment long video sequences effectively. Recent innovations continue to push the boundaries of memory utilization. S-ViT (Zhao et al., 2023) and MMC (Siam et al., 2023) focus on preserving detailed feature maps and reducing background confusion through multiscale memory transformers. RFGM (Zhou et al., 2023) introduces a relevance attention mechanism to adaptively assist in selecting pertinent historical information. RMem (Zhou et al., 2024) enhances efficiency by restricting memory banks to essential frames. MAVOS (Shaker et al., 2024) optimizes long-term memory usage to ensure temporal smoothness without frequent expansions, and SAM 2 (Ravi et al., 2024) extends memory capabilities with a FIFO queue for seamless object tracking.

Action level Task Action classification and localization in videos are critical components of understanding dynamic scenes and require sophisticated models capable of identifying and interpreting temporal patterns. Recent advancements in this field have leveraged memory networks and transformer-based architectures to enhance the accuracy and efficiency of these tasks. One approach that has gained prominence is the use of memory networks. For instance, (Yuan et al., 2019) introduces a novel framework that writes significant information into an external memory module while discarding irrelevant data. This selective memory management improves video action recognition by focusing on key temporal elements. Transformer-based models have also been instrumental in advancing action classification and localization. LSTR (Xu et al., 2021) utilizes both long-term and short-term memory in a FIFO structure to address online action detection, allowing the model to maintain relevant past information efficiently. MeMViT (Wu et al., 2022a) proposes a hierarchical memory transformer that optimizes long-term memory use for effective video action classification and anticipation. Another significant contribution is RViT (Yang et al., 2022), which employs an attention gate to facilitate interaction between the current frame input and the previous hidden state. This mechanism enhances the model's ability to integrate past and present information dynamically. Similarly, MAT (Wang et al., 2023b) introduces a memory encoder that compresses both long-term and short-term memory in a segment-based manner. It also features a memory-anticipation circular decoder that updates historical and future representations for online action detection and anticipation. Finally, MATR (Song et al., 2024) presents a FIFO memory queue that selectively retains past segment features, optimizing the process of temporal action localization. This approach ensures that the most relevant temporal features are preserved, improving the model's ability to localize actions accurately over time.

General level Task Video summarization has evolved significantly with the introduction of memory-augmented models and deep learning architectures. Early approaches, such as vsLSTM (Zhang et al., 2016), employed LSTM to capture variable-range temporal dependencies among video frames. This method aimed to create both representative and compact video summaries by effectively modeling the temporal relationships inherent in video data. Building on this foundation, MAVS (Feng et al., 2018) introduced a memory-augmented extractive video summarizer that utilized an external memory to store comprehensive visual information from the entire video, enhancing the summarization process with high-capacity memory storage. Furthermore, SMN (Wang et al., 2019) stacked multiple LSTM and memory layers hierarchically. This approach integrated learned representations from prior layers, resulting in more precise video summaries for individual frames by capturing intricate temporal patterns.

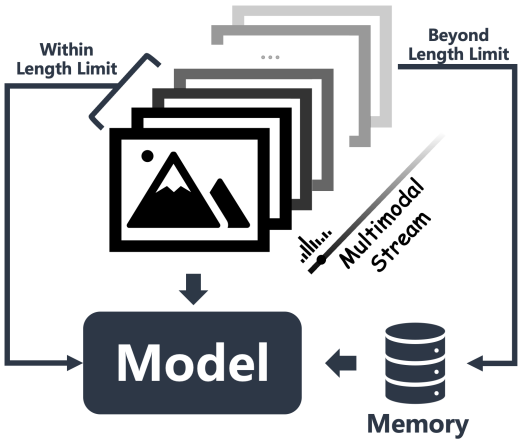

Figure 15: Memory for multimodal context modeling

Video captioning has also benefited from advancements in memory-augmented networks. M3 (Wang et al., 2018) proposed attaching an LSTM with an external memory that could store and retrieve both visual and textual content. This method allowed for multiple read and write operations, facilitating rich interactions between video sequences and corresponding sentences. MARN (Pei et al., 2019) introduced the Memory-Attended Recurrent Network, designed to explore the full-spectrum correspondence between words and their visual contexts, enhancing the captioning capabilities by leveraging memory structures. In terms of transformer-based solutions, MART (Lei et al., 2020) utilized a layer-wise TransformerXL architecture for video captioning, which was further improved by AAP-MIT (Prudviraj et al., 2022) through the integration of a Pyramid network for generating multi-sentence video descriptions, thereby enriching the narrative depth of the captions.

Video QA tasks have seen significant enhancements through the incorporation of sophisticated memory mechanisms and attention models. The Heterogeneous Memory Enhanced Multimodal Attention Model (Fan et al., 2019) introduced a heterogeneous external memory module on LSTM. This model employed attentional read and write operations to integrate motion and appearance features, co-learning attention mechanisms, and utilizing visual-question interactions to derive global context-aware representations. (Cai et al., 2020) further advanced Video QA by introducing fine-grained feature-augmented memories. This approach strengthened the information augmentation of video and text, improving memory capacity by capturing global interactions between high-level semantic information through self-attention and co-attention modules. RWMN (Na et al., 2017) utilized multi-layered CNNs to read and write sequential memory cells as chunks, effectively representing sequential stories with strong inter-block correlations. Lastly, Glance-Focus (Bai et al., 2023) proposed a two-stage method for Video QA. In the "glance" stage, an Encoder-Decoder generated dynamic event memories without supervision, while in the "focus" stage, these memories bridged the correlation between questions and both high-level event concepts and low-level video content, enhancing the model's comprehension and response accuracy.

Other Understaning Task In addition to common video understanding tasks, memory-augmented models have been increasingly applied to various other video processing tasks, including flow estimation, depth estimation, video deblurring, gesture recognition, and visual speech recognition. Flow estimation has benefited from models like MemFlow (Dong & Fu, 2024), which employs memory storage for real-time flow estimation, retaining motion information to enhance accuracy. In depth estimation, MAMo (Yasarla et al., 2023) augments networks with memory to store learned visual and displacement tokens from previous frames, allowing for more accurate depth predictions through cross-referencing past features. Video deblurring has advanced with MmDeblur (Ji & Yao, 2022), which uses a memory branch to memorize blurry-sharp feature pairs, aiding the deblurring process for incoming frames. Gesture recognition has been improved by MENet (Sun et al., 2022), which features a dual-branch architecture to capture temporal dynamics between spatiotemporal windows. Visual speech recognition has seen enhancements through frameworks using associative bridges to

learn interrelationships and obtain target modal representations from memory (Kim et al., 2021), with VAM (Kim et al., 2022a) imprinting audio features into a memory network using visual features, and MVM (Kim et al., 2022b) employing multihead key memories for visual features and a value memory for audio knowledge to distinguish homophenes. These applications show the versatility of memory networks in enhancing video understanding systems by leveraging past information for improved accuracy and robustness in complex video analysis scenarios.

Generation Task In the realm of video generation, there is a growing demand for creating long, high-quality videos. This challenge has led to the development of various memory-augmented models that enhance the quality and coherence of generated video content by leveraging advanced memory.

In the realm of video generation, there is a growing demand for creating long, high-quality videos, leading to the development of various memory-augmented models that enhance video content by leveraging advanced memory networks. The LMC-Memory model (Lee et al., 2021) utilizes memory alignment learning to store long-term motion contexts, improving video prediction by matching these contexts with sequences that exhibit limited dynamics. Similarly, MV-TON (Zhong et al., 2021) introduces a memory refinement module that embeds generated frames into a latent space as external memory, aiding subsequent frame generation with richer context. For talking face video generation, MemGAN (Yi et al., 2020) incorporates a memory-augmented GAN module to refine roughly rendered frames into realistic ones, enhancing video quality. Building on this, SyncTalkFace (Park et al., 2022) introduces an Audio-Lip Memory mechanism to align visual information of the mouth region with input audio, ensuring fine-grained audio-visual coherence. The EMMN model (Tan et al., 2023) constructs a Motion Memory Net that stores emotion embeddings and mouth motion features as key-value pairs, ensuring consistency between expression and lip motion. STAM (Chang et al., 2023) enhances spatiotemporal memorizing capacity by using a SpatioTemporal Attention based Memory on 3D-CNN, incorporating global spatiotemporal information to improve video prediction. Lastly, MemFace (Tang et al., 2022) addresses missing information in video generation with implicit and explicit memory components, capturing high-level semantics in the audio-expression shared space and aiding the neural-rendering model in synthesizing pixel-level details. These innovations underscore the critical role of memory networks in advancing video generation technology, enabling coherent and visually appealing outputs by effectively integrating past information and context.

5.3 Multimodal Contextual Memory for Robotics

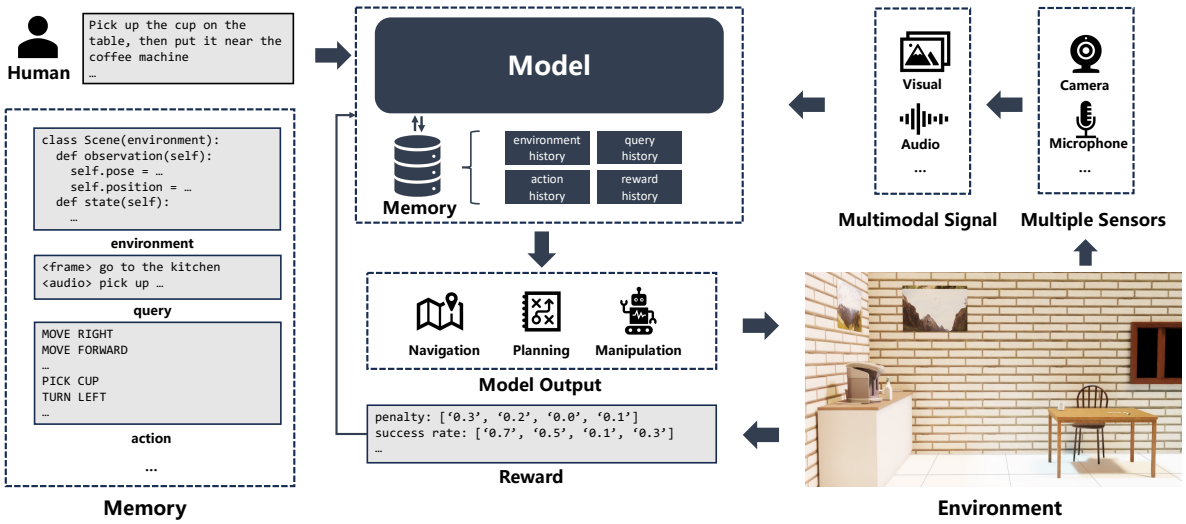

Figure 16: Memory for robotics

Integrating memory mechanisms into the capabilities of embodied agents and robotics has become increasingly essential for enhancing long-term planning, decision-making, visual navigation, and manipulation reasoning. These advancements demonstrate how memory can significantly improve an agent's ability to operate in complex environments.

5.3.1 Multimodal Memory-Augmented Agents

Multimodal memory-augmented agents demonstrate the integration of memory in diverse environments. JARVIS-1 (Wang et al., 2023j) equips an agent to perceive multimodal inputs, generate complex plans, and perform embodied control in a gaming environment like Minecraft, using memory to combine pre-trained knowledge with actual game experiences. MM-Navigator (Yan et al., 2023) employs GPT-4V for multimodal self-summarization in smartphone GUI navigation tasks, converting historical actions into concise natural language memory. MobileGPT (Lee et al., 2023) and AppAgent (Zhang et al., 2023b) focus on smartphone applications, accumulating knowledge about apps in graph form and summarizing interaction histories for improved decision-making and interpretability. OS-Copilot (Wu et al., 2024) features a framework for building generalist agents capable of interacting with various elements in an operating system, using a configurator with working, declarative, and procedural memory. MEIA (Liu et al., 2024a) offers an embodied agent for cafe scenes, utilizing a multimodal environment memory module that stores key scene information in natural language, guiding large models to execute action plans effectively under diverse requirements.

5.3.2 Memory-Enhanced Navigation, Odometry, and Manipulation

In the field of visual navigation, memory mechanisms play a crucial role in enabling agents to navigate effectively. RPF (Kumar et al., 2018) abstracts sequences of images and actions into memories for robust path following using RNNs. SSM (Wang et al., 2021a) introduces an external structured memory that stores visual and geometric information in disentangled layouts, providing a global action space on LSTM for visual navigation. VGM (Kwon et al., 2021) presents visual graph memory based on GCN, which includes unsupervised image representations for navigation history. In the realm of image-goal navigation, memory-augmented reinforcement learning (Mezghani et al., 2022) integrates an external memory mechanism with representations of past observations into the navigation policy. MemoNav (Li et al., 2024a) introduces a memory model based on GCN and LSTM, attending to short- and long-term memory while efficiently managing memory by forgetting information below a threshold. SMT (Fang et al., 2019) incorporates attention mechanisms to exploit spatio-temporal dependencies, maintaining long time horizons for navigation. Furthermore, MultiON (Wani et al., 2020) proposes navigation tasks to test agents' ability to locate previously observed goal objects. In visual odometry, memory mechanisms also enhance performance. The Deep Visual Odometry With Adaptive Memory model (Xue et al., 2022) employs selective memory based on RNNs to improve visual odometry accuracy and adaptability. For manipulation reasoning and planning, RDMemory (Huang et al., 2024b) encodes object-oriented memory into a multi-object manipulation framework based on transformers, facilitating sophisticated reasoning and planning capabilities.

5.3.3 Application

Memory is a critical component in the evolution of multimodal embodied agents, enabling them to seamlessly integrate and process diverse inputs such as visual, auditory, and textual data for adaptive, context-aware decision-making. Recent advancements highlight the role of memory-enhanced agents in tasks like autonomous navigation, healthcare assistance, interactive education, and smart home systems. By leveraging memory, these agents can retain past interactions, learn from experiences, and adapt to complex, dynamic environments, significantly enhancing their ability to understand, plan, and execute tasks across domains. Applications range from disaster response robots that utilize spatial memory for efficient navigation to personalized assistants that adapt based on user preferences and history. Memory's role in providing continuity and context allows these agents to go beyond static task execution, achieving higher levels of intelligence and functionality in both real-world and virtual scenarios.

5.4 Limitations and Future Works

While existing research has achieved significant progress in long-sequence multimodal tasks—such as long video understanding and long document processing—horizontal scaling challenges remain particularly pronounced in multimodal systems. This stems from the inherent abundance of visual tokens generated by patch-based image processing methods. Two critical factors exacerbate these challenges:

- Multimodal interaction inherently demands multi-turn reasoning, necessitating robust long-term memory to retain contextual coherence.
- Time-series modalities (e.g., audio, video, or streaming data) require long-term memory retention to model temporal dependencies effectively.
- Embodied learning requires memorizing multimodal information from the interaction between the agent and the real world.

Addressing these memory challenges—balancing computational efficiency with model effectiveness—represents a pivotal frontier for advancing multimodal systems.

6 Conclusion

In this report, we present a narrative review of three distinct types of memory integrated into large language models (LLMs): implicit memory, which is embedded within model parameters; explicit memory, which involves external storage and retrieval mechanisms; and agent memory, which captures persistent interactions with environments. Additionally, we systematically examine memory mechanisms specifically designed for and utilized by multimodal LLMs.

Based on this comprehensive review, we outline several key considerations and future research directions. First, there is a critical need to advance our understanding of the internal mechanisms of Transformer architectures and to develop more effective frameworks for implicit memory modeling. Second, enhancing the long-context processing capabilities of LLMs, either through extended context windows or retrieval-augmented generation (RAG), is essential; however, each approach presents trade-offs in terms of computational efficiency and scalability. Third, dynamic memory adaptation, inspired by human learning strategies such as recursive retrieval and experience reflection, holds promise for improving reasoning and communication in agent-based systems. Finally, multimodal systems face particular challenges stemming from the high volume of visual tokens, the complexity of multi-turn reasoning, and the temporal dependencies inherent in time-series data. Addressing these challenges calls for the development of scalable, memory-efficient architectures that support coherent, adaptive, and multimodal long-term learning.

We invite fellow researchers to engage with our findings and collaborate toward the advancement of memory-augmented AI systems.

References

Aadharsh Aadhithya A et al. Enhancing long-term memory using hierarchical aggregate tree for retrieval augmented generation. *arXiv e-prints*, pp. arXiv–2406, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning (ICML)*, 2024b.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Richard C Atkinson. Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2, 1968.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong Park, and Sung Hwang. Knowledge-augmented language model verification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1720–1736, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.107. URL https://aclanthology.org/2023.emnlp-main.107.

Ziyi Bai, Ruiping Wang, and Xilin Chen. Glance and focus: Memory prompting for multi-event video question answering. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J. Hénaff. Memory consolidation enables long-context video understanding. *CoRR*, abs/2402.05861, 2024.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. Unlimiformer: Long-range transformers with unlimited length input. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35522–35543. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6f9806a5adc72b5b834b27e4c7c0df9b-Paper-Conference.pdf.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: a memory viewpoint. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, pp. 2206–2240. PMLR, 2022.

Aljaz Bozic, Pablo R. Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular RGB scene reconstruction using transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1403–1414, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiayin Cai, Chun Yuan, Cheng Shi, Lei Li, Yangyang Cheng, and Ying Shan. Feature augmented memory with global attention network for videoqa. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 998–1004. ijcai.org, 2020.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Carlos Carvalho and Alberto Abad. Memory-augmented conformer for improved end-to-end long-form ASR. In Naomi Harte, Julie Carson-Berndsen, and Gareth Jones (eds.), *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 2218–2222. ISCA, 2023.

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. STAM: A spatiotemporal attention based memory for video prediction. *IEEE Trans. Multim.*, 25:2354–2367, 2023.

Harrison Chase. LangChain, October 2022. URL https://github.com/langchain-ai/langchain.

Anwoy Chatterjee, Eshaan Tanwar, Subhabrata Dutta, and Tanmoy Chakraborty. Language models can exploit cross-task in-context learning for data-scarce novel tasks. *arXiv preprint arXiv:2405.10548*, 2024.

Subhajit Chaudhury, Payel Das, Sarathkrishna Swaminathan, Georgios Kollias, Elliot Nelson, Khushbu Pahwa, Tejaswini Pedapati, Igor Melnyk, and Matthew Riemer. Epman: Episodic memory attention for generalizing to longer contexts. *arXiv preprint arXiv:2502.14280*, 2025.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. Understanding retrieval augmentation for long-form question answering. *ArXiv*, abs/2310.12150, 2023. URL https://api.semanticscholar.org/CorpusID:264288955.

Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023.

Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations, 2024a. URL https://arxiv.org/abs/2402.11975.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*, 2024b.

Zeming Chen, Gail Weiss, Eric Mitchell, Asli Celikyilmaz, and Antoine Bosselut. Reckoning: reasoning through dynamic knowledge encoding. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024c.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12318–12337, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.758. URL https://aclanthology.org/2023.emnlp-main.758.

Ho Kei Cheng and Alexander G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pp. 640–658. Springer, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11781–11794, 2021.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Lawrence Chisvin and R. James Duckworth. Content-addressable and associative memory**based on "content-addressable and associative memory: Alternatives to the ubiquitous ram" by lawrence chisvin and r. james duckworth which appeared in ieee computer, vol. 22, no. 7, pages 51–64, july 1989. copyright © 1989 ieee. volume 34 of *Advances in Computers*, pp. 159–235. Elsevier, 1992. doi: https://doi.org/10.1016/S0065-2458(08)60326-5.

Kevin Clark. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Trans. Assoc. Comput. Linguistics*, 2024.

Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. SIGIR '24, pp. 719–729. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657834. URL https://doi.org/10.1145/3626772.3657834.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9465–9480, 2022.

Michiel De Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William Cohen. Mention memory: incorporating textual knowledge into transformers through entity mention attention. *arXiv preprint arXiv:2110.06176*, 2021.

Ilana TZ Dew and Roberto Cabeza. The porous boundaries between explicit and implicit memory: behavioral and neural evidence. *Annals of the New York Academy of Sciences*, 1224(1):174–190, 2011.

Dai Do, Quan Tran, Svetha Venkatesh, and Hung Le. Large language models prompting with episodic memory. *ArXiv*, abs/2408.07465, 2024. URL https://api.semanticscholar.org/CorpusID:271865662.

Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. *CoRR*, abs/2404.04808, 2024.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 2022.

Mingzhe Du, Anh Tuan Luu, Bin Ji, and See-kiong Ng. From static to dynamic: A continual learning framework for large language models. *arXiv preprint arXiv:2310.14248*, 2023.

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

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1999–2007. Computer Vision Foundation / IEEE, 2019.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *CoRR*, abs/2403.11481, 2024.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 538–547. Computer Vision Foundation / IEEE, 2019.

Shiyu Fang, Jiaqi Liu, Chengkai Xu, Chen Lv, Peng Hang, and Jian Sun. Interact, instruct to improve: A llm-driven parallel actor-reasoner framework for enhancing autonomous vehicle interactions. 2025. URL https://api.semanticscholar.org/CorpusID:276742216.

Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. Extractive video summarizer with memory augmented neural networks. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (eds.), *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pp. 976–983. ACM, 2018.

Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 13774–13783. Computer Vision Foundation / IEEE, 2021.

Hang Gao and Yongfeng Zhang. Memory sharing for large language model based agents. *arXiv preprint arXiv:2404.09982*, 2024.

Manas Gaur, Kalpa Gunaratna, Vijay Srinivasan, and Hongxia Jin. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. *ArXiv*, abs/2112.07622, 2021. URL https://api.semanticscholar.org/CorpusID:245131215.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022b.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jing Guo, Nan Li, Jianchuan Qi, Hang Yang, Ruiqiao Li, Yuzhen Feng, Si Zhang, and Ming Xu. Empowering working memory for large language model agents. *arXiv preprint arXiv:2312.17259*, 2023a.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. *ArXiv*, abs/2305.17653, 2023b. URL https://api.semanticscholar.org/CorpusID:258960340.

Anshita Gupta, Debanjan Mondal, Akshay Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. Editing common sense in transformers. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Priyanshu Gupta, Shashank Kirtania, Ananya Singha, Sumit Gulwani, Arjun Radhakrishna, Sherry Shi, and Gustavo Soares. Metareflection: Learning instructions for language agents using past reflections. *arXiv preprint arXiv:2405.13009*, 2024.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. Demix layers: Disentangling domains for modular language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics, 2022.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, pp. 3929–3938. PMLR, 2020.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. *arXiv preprint arXiv:2109.06780*, 2021.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2714–2731, 2023.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: memory-augmented large multimodal model for long-term video understanding. *CoRR*, abs/2404.05726, 2024a.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*, 2024b.

Zexue He, Leonid Karlinsky, Donghyun Kim, Julian McAuley, Dmitry Krotov, and Rogerio Feris. Camelot: Towards large language models with training-free consolidated associative memory, 2024c.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exbert: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8771–8780. IEEE, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 7732–7741. Computer Vision Foundation / IEEE, 2021.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. Pip-kag: Mitigating knowledge conflicts in knowledge-augmented generation via parametric pruning. *arXiv preprint arXiv:2502.15543*, 2025.

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

Yixuan Huang, Jialin Yuan, Chanho Kim, Pupul Pradhan, Bryan Chen, Fuxin Li, and Tucker Hermans. Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 3108–3115. IEEE, 2024b.

Wenyang Hui, Yan Wang, Kewei Tu, and Chengyue Jiang. Rot: Enhancing large language models with reflection on search trees. *arXiv preprint arXiv:2404.05449*, 2024.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research (JMLR)*, 24(251):1–43, 2023.

Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6558–6567. Computer Vision Foundation / IEEE, 2021.

Bo Ji and Angela Yao. Multi-scale memory-based video deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 1909–1918. IEEE, 2022.

Zixia Jia, Mengmeng Wang, Baichen Tong, Song-Chun Zhu, and Zilong Zheng. Langsuite: Planning, controlling and interacting with large language models in embodied text environments. *arXiv preprint arXiv:2406.16294*, 2024.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. 2023. URL https://api.semanticscholar.org/CorpusID:266551228.

Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, et al. Long term memory: The foundation of ai self-evolution. *arXiv preprint arXiv:2410.15665*, 2024a.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers, 2024b. URL https://arxiv.org/abs/2406.18400.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, and Bryon Aragam. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024c.

Youngsaeng Jin, David K. Han, and Hanseok Ko. Memory-based semantic segmentation for off-road unstructured natural environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pp. 24–31. IEEE, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

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

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. Language repository for long video understanding. *CoRR*, abs/2403.14622, 2024.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. Cross-lingual multi-hop knowledge editing–benchmarks, analysis and a simple contrastive learning based approach. *arXiv preprint arXiv:2407.10275*, 2024.

Savya Khosla, Zhen Zhu, and Yifei He. Survey on memory-augmented neural networks: Cognitive insights to ai applications. *arXiv preprint arXiv:2312.06141*, 2023.

Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 296–306. IEEE, 2021.

Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Cromm-vsr: Cross-modal memory augmented visual speech recognition. *IEEE Trans. Multim.*, 24:4342–4355, 2022a.

Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 1174–1182. AAAI Press, 2022b.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

Dmitry Krotov. Hierarchical associative memory. 2021. URL https://arxiv.org/abs/2107.06446.

Ashish Kumar, Saurabh Gupta, David F. Fouhey, Sergey Levine, and Jitendra Malik. Visual memory for robust path following. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 773–782, 2018.

Obin Kwon, Nuri Kim, Yunho Choi, Hwiyeon Yoo, Jeongho Park, and Songhwai Oh. Visual graph memory with unsupervised representation for visual navigation. In *International Conference on Computer Vision (ICCV)*, pp. 15870–15879. IEEE, 2021.

Jakub L'ala, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research. *ArXiv*, abs/2312.07559, 2023. URL https://api.semanticscholar.org/CorpusID:266191420.

Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. Copy is all you need. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=CROlOA9Nd8C.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 3054–3063. Computer Vision Foundation / IEEE, 2021.

Sunjae Lee, Junyoung Choi, Jungjae Lee, Hojun Choi, Steven Y. Ko, Sangeun Oh, and Insik Shin. Explore, select, derive, and recall: Augmenting LLM with human-like memory for mobile task automation. *CoRR*, abs/2312.03003, 2023.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2603–2614. Association for Computational Linguistics, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, and Zilong Zheng. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space, 2025a. URL https://arxiv.org/abs/2505.13308.

Hongxin Li, Zeyu Wang, Xu Yang, Yuran Yang, Shuqi Mei, and Zhaoxiang Zhang. Memonav: Working memory model for visual navigation. *CoRR*, abs/2402.19161, 2024a.

Jia Li, Huaibo Huang, Xiaofei Jia, and Ran He. Universal face restoration with memorized modulation. *CoRR*, abs/2110.01033, 2021.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023a.

Jiaqi Li, Xiaobo Wang, Wentao Ding, Zihao Wang, Yipeng Kang, Zixia Jia, and Zilong Zheng. Ram: Towards an ever-improving memory system by learning from communications. *arXiv preprint arXiv:2404.12045*, 2024b.

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, SongChun Zhu, Zixia Jia, and Zilong Zheng. Reflectevo: Improving meta introspection of small llms by learning self-reflection. *arXiv preprint arXiv:2505.16475*, 2025b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024d.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. Structure-aware language model pretraining improves dense retrieval on structured data. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11560–11574, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.734. URL https://aclanthology.org/2023.findings-acl.734.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke S. Zettlemoyer, and Scott Yih. Ra-dit: Retrieval-augmented dual instruction tuning. *ArXiv*, abs/2310.01352, 2023. URL https://api.semanticscholar.org/CorpusID:263605962.

Jerry Liu. LlamaIndex, 11 2022. URL https://github.com/jerryjliu/llama_index.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023.

Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. Multimodal embodied interactive agent for cafe scene. *CoRR*, abs/2402.00290, 2024a.

Yuhan Liu, Esha Choukse, Shan Lu, Junchen Jiang, and Madan Musuvathi. Droidspeak: Enhancing cross-llm communication. *arXiv preprint arXiv:2411.02820*, 2024b.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024c.

Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pp. 661–679. Springer, 2020.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact memorization of large language models. *arXiv preprint arXiv:2406.15720*, 2024.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. *ArXiv*, abs/2310.01061, 2023. URL https://api.semanticscholar.org/CorpusID:263605944.

Ang Lv, Kaiyi Zhang, Yuhan Chen, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. Interpreting key mechanisms of factual recall in transformer-based language models. *Association for Computing Machinery (ACM*, 2024.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *ArXiv*, abs/2305.14283, 2023. URL https://api.semanticscholar.org/CorpusID:258841283.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. *CoRR*, abs/2406.12846, 2024.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve GPT-3 after deployment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2833–2861, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.183. URL https://aclanthology.org/2022.emnlp-main.183.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*, 2022b.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

Jinjie Mai, Jun Chen, Guocheng Qian, Mohamed Elhoseiny, Bernard Ghanem, et al. Llm as a robotic brain: Unifying egocentric memory and control. 2023.

Yihuan Mao, Yipeng Kang, Peilun Li, Ning Zhang, Wei Xu, and Chongjie Zhang. Ibgp: Imperfect byzantine generals problem for zero-shot robustness in communicative multi-agent systems. *arXiv preprint arXiv:2410.16237*, 2024.

Samuele Marro, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael Wooldridge, and Philip Torr. A scalable communication protocol for networks of large language models. *arXiv preprint arXiv:2410.11905*, 2024.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Eric Melz. Enhancing llm intelligence with arm-rag: Auxiliary rationale memory for retrieval augmented generation. *arXiv preprint arXiv:2311.04177*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 3316–3323. IEEE, 2022.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022, Baltimore, Maryland, USA, 17-23 July 2022*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15561–15583. PMLR, 2022. URL https://proceedings.mlr.press/v162/millidge22a.html.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2023.

Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 677–685. IEEE Computer Society, 2017.

Xueyan Niu, Bo Bai, Lei Deng, and Wei Han. Beyond scaling laws: Understanding transformer performance with associative memory. *CoRR*, abs/2405.08707, 2024. URL https://doi.org/10.48550/arXiv.2405.08707.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9225–9234. IEEE, 2019.

Shankar Padmanabhan, Yasumasa Onoe, Michael J. Q. Zhang, Greg Durrett, and Eunsol Choi. Propagating knowledge updates to lms through distillation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6432–6440. IEEE Computer Society, 2017.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 2062–2070. AAAI Press, 2022.

Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pp. 1102–1109. IEEE, 2021.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8347–8356. Computer Vision Foundation / IEEE, 2019.

Jeripothula Prudviraj, Malipatel Indrakaran Reddy, Chalavadi Vishnu, and Chalavadi Krishna Mohan. AAP-MIT: attentive atrous pyramid network and memory incorporated transformer for multisentence video description. *IEEE Trans. Image Process.*, 31:5559–5569, 2022.

Shi Pu, Yibing Song, Chao Ma, Honggang Zhang, and Ming-Hsuan Yang. Learning recurrent memory activation networks for visual tracking. *IEEE Trans. Image Process.*, 30:725–738, 2021.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *CoRR*, abs/2405.16009, 2024.

Zihan Qiu, Zeyu Huang, Youcheng Huang, and Jie Fu. Empirical study on updating key-value memories in transformer feed-forward layers. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2021.

Priyanka Ranade and Anupam Joshi. Fabula: Intelligence report generation using retrieval-augmented narrative construction. *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023. URL https://api.semanticscholar.org/CorpusID:264425965.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019, 2023a. URL https://api.semanticscholar.org/CorpusID:259991467.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 293–306, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.17. URL https://aclanthology.org/2023.acl-long.17.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. *ArXiv*, abs/2401.18059, 2024. URL https://api.semanticscholar.org/CorpusID:267334785.

Abdelrahman M. Shaker, Syed Talal Wasim, Martin Danelljan, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Efficient video object segmentation via modulated cross-attention memory. *CoRR*, abs/2403.17937, 2024.

Lianlei Shan, Shixian Luo, Zezhou Zhu, Yu Yuan, and Yong Wu. Cognitive memory in large language models. *arXiv preprint arXiv:2504.02441*, 2025.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9248–9274, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.620. URL https://aclanthology.org/2023.findings-emnlp.620.

Aditya Sharma, Luis Lara, Amal Zouaq, and Christopher J Pal. Reducing hallucinations in language model-based sparql query generation using post-generation memory retrieval. *arXiv preprint arXiv:2502.13369*, 2025.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Mennatullah Siam, Rezaul Karim, He Zhao, and Richard Wildes. Multiscale memory comparator transformer for few-shot video segmentation. *CoRR*, abs/2307.07812, 2023.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *CoRR*, abs/2307.16449, 2023.

Youngkil Song, Dongkeun Kim, Minsu Cho, and Suha Kwak. Online temporal action localization with memory-augmented transformer, 2024.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Haitian Sun, Patrick Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, and William W Cohen. Reasoning over virtual knowledge bases with open predicate relations. In *International Conference on Machine Learning*, pp. 9966–9977. PMLR, 2021.

Linhui Sun, Yifan Zhang, Ke Cheng, Jian Cheng, and Hanqing Lu. Menet: A memory-based network with dual-branch for efficient event stream processing. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, volume 13684 of *Lecture Notes in Computer Science*, pp. 214–234. Springer, 2022.

Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, and Feng Wu. Alignment before aggregation: Trajectory memory retrieval network for video object segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 1218–1228. IEEE, 2023.

Wendy A. Suzuki. Chapter 19 associative learning signals in the brain. In Wayne S. Sossin, Jean-Claude Lacaille, Vincent F. Castellucci, and Sylvie Belleville (eds.), *Essence of Memory*, volume 169 of *Progress in Brain Research*, pp. 305–320. Elsevier, 2008. doi: https://doi.org/10.1016/S0079-6123(07)00019-2.

Hao Hao Tan, Kin Wai Cheuk, Taemin Cho, Wei-Hsiang Liao, and Yuki Mitsufuji. MR-MT3: memory retaining multi-track music transcription to mitigate instrument leakage. *CoRR*, abs/2403.10024, 2024.

Shuai Tan, Bin Ji, and Ye Pan. EMMN: emotional motion memory network for audio-driven emotional talking face generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 22089–22099. IEEE, 2023.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *arXiv preprint arXiv:2112.09737*, 2021.

Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. Memories are one-to-many mapping alleviators in talking face generation. *CoRR*, abs/2212.05005, 2022.

Xiaojuan Tang, Jiaqi Li, Yitao Liang, Song-Chun Zhu, Muhan Zhang, and Zilong Zheng. Mars: Situated inductive reasoning in an open-world environment. *Advances in Neural Information Processing Systems*, 37:17830–17869, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*, 2024.

Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 4491–4500. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.480. URL https://doi.org/10.1109/ICCV.2017.480.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.

Szymon Tworkowski, Konrad Staniszewski, Mikoł aj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Mił oś. Focused transformer: Contrastive training for context scaling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 42661–42688. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8511d06d5590f4bda24d42087802cc81-Paper-Conference.pdf.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Akshaj Kumar Veldanda, Shi-Xiong Zhang, Anirban Das, Supriyo Chakraborty, Stephen Rawls, Sambit Sahu, and Milind Naphade. Llm surgery: Efficient knowledge unlearning and editing in large language models. 2024. URL https://arxiv.org/abs/2409.13054.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia, 2023.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7763–7786, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.482. URL https://aclanthology.org/2023.emnlp-main.482.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. Instructretro: Instruction tuning post retrieval-augmented pretraining. In *International Conference on Machine Learning (ICML)*, 2024a. URL https://openreview.net/forum?id=PLAGGbssT8.

Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured scene memory for vision-language navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8455–8464. Computer Vision Foundation / IEEE, 2021a.

Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 13778–13789. IEEE, 2023b.

Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: multimodal memory modelling for video captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7512–7520. Computer Vision Foundation / IEEE Computer Society, 2018.

Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. Stacked memory network for video summarization. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (eds.), *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pp. 836–844. ACM, 2019.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *CoRR*, abs/2304.14407, 2023c.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*. Association for Computational Linguistics, 2021b.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning, 2024b.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*, 2023d.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74530–74543. Curran Associates, Inc., 2023e. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ebd82705f44793b6f9ade5a669d0f0bf-Paper-Conference.pdf.

Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 7055–7064. IEEE, 2022a.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Editing conceptual knowledge for large language models. *arXiv preprint arXiv:2403.06259*, 2024c.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *CoRR*, abs/2403.10517, 2024d.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. R^{3}Mem: Bridging memory retention and retrieval via reversible compression. *arXiv preprint arXiv:2502.15957*, 2025.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023f.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023g.

Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in egocentric videos. *CoRR*, abs/2312.05269, 2023h.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. Large scale knowledge washing. *arXiv preprint arXiv:2405.16720*, 2024e.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges. *arXiv preprint arXiv:2409.01071*, 2024f.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation. *ArXiv*, abs/2311.08377, 2023i. URL https://api.semanticscholar.org/CorpusID:265157538.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. JARVIS-1: open-world multi-task agents with memory-augmented multimodal language models. *CoRR*, abs/2311.05997, 2023j.

Saim Wani, Shivansh Patel, Unnat Jain, Angel X. Chang, and Manolis Savva. Multion: Benchmarking semantic map memory using multi-object navigation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 13577–13587. IEEE, 2022a.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*, 2023a.

Tong Wu, Junzhe Shen, Zixia Jia, Yuxuan Wang, and Zilong Zheng. From hours to minutes: Lossless acceleration of ultra long sequence generation up to 100k tokens. *arXiv preprint arXiv:2502.18890*, 2025a.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. Depn: Detecting and editing privacy neurons in pretrained language models. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

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

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=TrjbxzRcnf-.

Yuqi Wu, Guangya Wan, Jingjing Li, Shengming Zhao, Lingfeng Ma, Tianyi Ye, Ion Pop, Yanbo Zhang, and Jie Chen. Proai: Proactive multi-agent conversational ai with structured knowledge base for psychiatric diagnosis. *arXiv preprint arXiv:2502.20689*, 2025c.

Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. Os-copilot: Towards generalist computer agents with self-improvement. *CoRR*, abs/2402.07456, 2024.

Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 497–506. IEEE, 2021.

Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1086–1099, 2021.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence C. McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *ArXiv*, abs/2310.03025, 2023. URL https://api.semanticscholar.org/CorpusID:263620134.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Fei Xue, Xin Wang, Junqiu Wang, and Hongbin Zha. Deep visual odometry with adaptive memory. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(2):940–954, 2022.

An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian J. McAuley, Jianfeng Gao, Zicheng Liu, and Lijuan Wang. GPT-4V in wonderland: Large multimodal models for zero-shot smartphone GUI navigation. *CoRR*, abs/2311.07562, 2023.

Hanqi Yan, Qinglin Zhu, Xinyu Wang, Lin Gui, and Yulan He. Mirror: A multiple-perspective self-reflection method for knowledge-rich reasoning. *arXiv preprint arXiv:2402.14963*, 2024.

Hanqing Yang, Jingdi Chen, Marie Siew, Tania Lorido-Botran, and Carlee Joe-Wong. Llm-powered decentralized generative agents with adaptive hierarchical knowledge graph for cooperative planning. *arXiv preprint arXiv:2502.05453*, 2025.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5364–5375, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.326. URL https://aclanthology.org/2023.emnlp-main.326.

Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, et al. *Memory*^{3}: Language modeling with explicit memory. *arXiv preprint arXiv:2407.01178*, 2024a.

Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14043–14053. IEEE, 2022.

Tianyu Yang and Antoni B. Chan. Recurrent filter learning for visual tracking. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pp. 2010–2019. IEEE Computer Society, 2017.

Tianyu Yang and Antoni B. Chan. Learning dynamic memory networks for object tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, volume 11213 of *Lecture Notes in Computer Science*, pp. 153–169. Springer, 2018.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2491–2502, 2021.

Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *CoRR*, abs/2401.08392, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Association for Computing Machinery (ACM*, 2024b.

Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 8720–8730. IEEE, 2023.

Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *ArXiv*, abs/2002.10137, 2020.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373, 2021.

Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11283–11292. Computer Vision Foundation / IEEE, 2019.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023a.

W. Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *ArXiv*, abs/2311.09210, 2023b. URL https://api.semanticscholar.org/CorpusID:265212816.

Yuan Yuan, Dong Wang, and Qi Wang. Memory-augmented temporal dynamic learning for action recognition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 9167–9175. AAAI Press, 2019.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. *CoRR*, abs/2312.17235, 2023a.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *CoRR*, abs/2312.13771, 2023b.

Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. Large language models are semi-parametric reinforcement learning agents. *Advances in Neural Information Processing Systems*, 36, 2024a.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *CoRR*, abs/2406.08085, 2024b.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023c.

Kaihua Zhang, Long Wang, Dong Liu, Bo Liu, Qingshan Liu, and Zhu Li. Dual temporal memory network for efficient video object segmentation. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (eds.), *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pp. 1515–1523. ACM, 2020.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pp. 766–782. Springer, 2016.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024c.

Siyuan Zhang, Yichi Zhang, Yinpeng Dong, and Hang Su. Self-memory alignment: Mitigating factual hallucinations with generalized improvement. *arXiv preprint arXiv:2502.19127*, 2025.

Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. Loop copilot: Conducting AI ensembles for music generation and iterative editing. *CoRR*, abs/2310.12404, 2023d.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024d.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. IAG: Induction-augmented generation framework for answering reasoning questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1–14, Singapore, December 2023e. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1. URL https://aclanthology.org/2023.emnlp-main.1.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. MoEfication: Transformer feed-forward layers are mixtures of experts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 877–890, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.71. URL https://aclanthology.org/2022.findings-acl.71.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics:*

ACL 2023, pp. 4066–4083, Toronto, Canada, July 2023f. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.250. URL https://aclanthology.org/2023.findings-acl.250.

Zhihan Zhang, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, and Meng Jiang. Learn beyond the answer: Training language models with reflection for mathematical reasoning. *arXiv preprint arXiv:2406.12050*, 2024e.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*, 2023g.

Zihan Zhang, Meng Fang, and Ling Chen. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *arXiv preprint arXiv:2402.16457*, 2024f.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. Streaming video model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14602–14612. IEEE, 2023.

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

Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. MV-TON: memory-based video virtual try-on network. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (eds.), *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 908–916. ACM, 2021.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. corr, abs/2305.14795, 2023. doi: 10.48550. *arXiv preprint arXiv.2305.14795*.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation. *arXiv preprint arXiv:2205.12674*, 2022a.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 2022b.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 2023.

Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. *CoRR*, abs/2406.08476, 2024.

Xinyu Zhou, Pinxue Guo, Lingyi Hong, Jinglun Li, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Reading relevant feature from global representation memory for visual object tracking. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*, 2020.