Agentic Episodic Control

Anonymous ACL submission

Abstract

Reinforcement learning (RL) remains fundamentally limited by poor data efficiency and weak generalization. Prior episodic RL methods attempt to alleviate this via external memory modules, yet suffer from two key limitations: a representation bottleneck caused by shallow encoders, and a retrieval dilemma where episodic memory is accessed indiscriminately. To address these challenges, we propose **Agentic Episodic Control** (AEC), a novel architecture that integrates large language models (LLMs) into episodic RL. AEC uses an LLM-based semantic augmenter to generate semantic representations from raw observations, and a critical state recognizer to selectively retrieve valuable experiences. This transforms memory usage from passive similarity matching into strategic, context-aware recall. Across five BabyAI-Text environments, AEC achieves 2–6× higher data efficiency than baselines and is the only method to solve complex tasks like UnlockLocal with over 90% success. It further demonstrates strong cross-task and cross-environment generalization, maintaining performance even under distribution shifts. AEC shows that combining LLM-derived priors with reinforcement learning yields more sample-efficient and adaptable agents.

1 Introduction

Reinforcement Learning (RL) has been a driving force in AI over the past decade (Silver et al., 2016; Tunyasuvunakool et al., 2021; Fawzi et al., 2022; Mankowitz et al., 2023). Today, RL remains pivotal through Reinforcement Learning from Human Feedback (RLHF) and reinforcement fine-tuning (RFT), which aligns large language models with human preferences (Ouyang et al., 2022; Trung et al., 2024). Despite these successes, fundamental challenges separate RL agents from human learners in two critical aspects: **data efficiency** and **generalization** (Schwarzer et al., 2021; Korkmaz, 2024). While humans learn robustly from a handful of experiences and seamlessly transfer knowledge, RL agents often require millions of interactions to master a single task and fail to adapt when faced with novel compositions. We argue that this gap stems from a deeper deficiency: *the absence of cognitive mechanisms that are central to human learning, particularly a memory system capable of semantic abstraction and strategic retrieval of relevant experiences* (Eichenbaum, 2004; Kuhl and Wagner, 2009).

Recognizing the importance of memory, prior works such as Neural Episodic Control (NEC) (Pritzel et al., 2017) and episodic reinforcement learning frameworks (Liang et al., 2024; Na et al., 2024) attempted to bridge this gap by incorporating external memory modules. These methods typically use a trainable neural network to generate state embeddings, which are stored in an episodic memory buffer. At each decision step, the agent performs a K-nearest-neighbor (KNN) lookup over this buffer to retrieve relevant past experiences—either using them directly to estimate Q-values for action selection, or to provide targets for updating the Q-network. However, because these designs rely on lightweight state encoders, they yield narrow, task-specific embeddings with insufficient semantic coverage for systematic generalization. Moreover, in sparse-reward settings, the episodic memory is dominated by near-zero-return trajectories, and per-timestep passive KNN reads therefore tend to retrieve low-value neighbors, which provide neither actionable signals nor useful training targets for the Q-network. Consequently, these pioneering efforts suffer from two critical flaws: a representation bottleneck, wherein shallow encoders produce retrieval keys with limited semantic and relational abstraction; and a retrieval dilemma, wherein the agent lacks the capacity to decide when or whether to retrieve from memory, defaulting instead to

unconditional KNN queries at each step.

Building on the above analysis, we argue that an effective framework still lacks an oracle—a module that, given the current state, infers latent task-relevant factors, completes the state representation accordingly, and decides whether to exploit memory or explore. Fortunately, we now stand at a unique inflection point, where the rise of large language models (LLMs) provides a practical path to the missing *oracle*. Trained on massive corpora that implicitly encode broad human knowledge, LLMs act as powerful semantic engines that support high-level reasoning—logical inference (Patel et al., 2024), commonsense reasoning (Wang and Zhao, 2023), and abstract problem solving (Imani et al., 2023). Crucially, by conditioning prompts on the current state and task context, we can elicit oracle-like signals: latent, task-relevant abstractions, candidate hypotheses, and decision heuristics that indicate whether to exploit prior experience or explore under distributional shift. In this sense, LLMs offer a readily accessible proxy for the oracle envisioned above.

This capability directly addresses the representation bottleneck by prompting LLMs with the current state and task context, allowing them to use their own commonsense knowledge to infer additional information beneath the surface state. Furthermore, their inherent capacity for logical inference and commonsense reasoning provides the foundation for building strategic, context-aware retrieval mechanisms, indicating whether to exploit prior experience or explore under a distributional shift, thus offering a clear path to resolving the retrieval dilemma.

Driven by the ability of LLMs, we introduce Agentic Episodic Control (AEC), a novel framework that addresses the representation bottleneck and retrieval dilemma by integrating two LLM-guided modules. An LLM-guided augmenter maps raw observations to language-grounded abstractions that expose latent, task-relevant structure. A critical-state recognizer then determines when consulting episodic memory is warranted and when to act using the agent's current understanding without recall. By turning memory access from passive similarity matching into selective, decision-aware recall, AEC yields improved data efficiency and generalization across varied task settings. Our contributions are four-fold:

- We identify two fundamental limitations in episodic RL: a representation bottleneck caused by shallow encoders, and a retrieval dilemma stemming from indiscriminate memory access.
- We propose Agentic Episodic Control (AEC), which addresses both limitations by leveraging large language models to generate abstract state representations and selectively retrieve memory based on task-critical context.
- Through experiments on five BabyAI-Text tasks, AEC achieves 2- to 6-fold improvements in data efficiency over baselines, and is the only method to solve complex tasks like UnlockLocal with near-perfect success.
- We further validate AEC's generalization across tasks and environments, showing robust performance under distribution shift and effective memory reuse across settings.

2 The AEC Framework

We propose an efficient episodic control framework called **Agentic Episodic Control** (see Figure 1), designed to simultaneously address the representation bottleneck and the retrieval dilemma of NEC and its variants. To overcome the representation bottleneck, AEC incorporates a semantic augmenter based on LLMs, which maps raw observations into language-aligned abstract representations. This yields richer feature embeddings and more stable keys for memory operations. To tackle the retrieval dilemma, AEC introduces a working memory that maintains a short-term summary of the current context to support exploration, along with a critical-state recognizer that dynamically mediates between exploitation and exploration. Episodic recall is selectively triggered only when it is likely to significantly influence future decisions; otherwise, the agent proceeds based on its current internal model. This transforms memory access from passive similarity-based retrieval into targeted, decision-aware recall, thereby enhancing data efficiency and generalization.

Preliminary. Episodic control draws on the hippocampus's memory mechanism (Eichenbaum, 2004). At each timestep t, the agent observes a state s_t. To enable memory-based decision-making, the agent employs a state embedding function $f_\phi(s) : \mathcal{S} \to \mathbb{R}^k$, which maps raw states into a k-dimensional latent space (Na et al., 2024). The resulting embedding is denoted by $z_t = f_\phi(s_t)$.

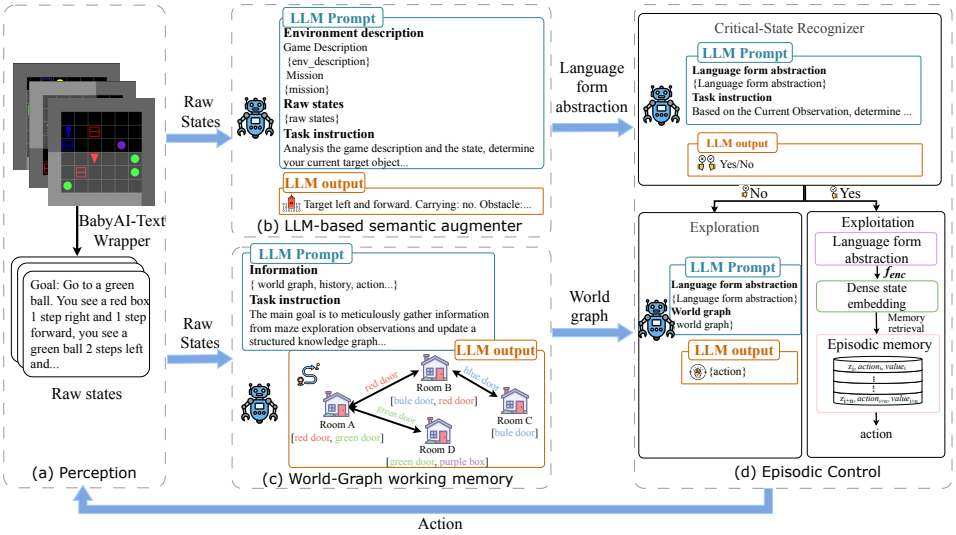

Figure 1: The overview of Agentic Episodic Control framework. Raw states are processed in parallel by an LLM-based semantic augmenter (b) to produce language-form abstraction and by a World-Graph working memory module (c) to build a structured working memory of entities and relations. An episodic control module (d), equipped with a critical-state recognizer, then uses these representations to decide whether to exploit episodic memory or to continue exploring under World-Graph guidance. The chosen actions interact with the environment.

For each action, the agent maintains an episodic memory buffer $\mathcal{D}_E^a$, which stores key–value pairs where each key is a state embedding z_t, and each value stores the highest return observed for a given state–action pair. Formally, given a pair (s_t, a_t), the agent updates the value $H(z_t)$ within the specific buffer $\mathcal{D}_E^{a_t}$ according to (Lin et al., 2018):

$$H(z_t) = \begin{cases} \max\{H(\hat{z}_t),\, R_t(s_t, a_t)\}, & \text{if } (s_t, a_t) \in \mathcal{D}_E^{a_t}, \\ R(s_t, a_t), & \text{otherwise,} \end{cases} \tag{1}$$

where $R(s_t, a_t)$ denotes the return received after taking action a_t in state s_t, and $\hat{z}_t$ represents the exact match of the current state–action pair in memory, if available.

At decision time, the agent queries the episodic memory for each candidate action a. Specifically, it performs a k-nearest neighbor (KNN) search over $\mathcal{D}_E^a$ using the current state embedding z_t, yielding a set of top-k closest entries $\mathcal{N}_p(z_t, \mathcal{D}_E^a)$. Each retrieved value is weighted according to its similarity to the query embedding, producing a kernel-weighted action value estimate:

$$Q(s_t, a) = \sum_{i \in \mathcal{N}_p(z_t, \mathcal{D}_E^a)} w_i \cdot H(z_i), \tag{2}$$

where the weights w_i are typically computed via a softmax over cosine similarities or a kernel function.

Finally, the selected action is:

$$a_t = \arg\max_a Q(s_t, a). \tag{3}$$

This memory-based approach enables rapid value estimation from past experience and underpins the enhanced episodic control mechanisms introduced in this work.

2.1 How to Solve the Representation Bottleneck?

To address the representation bottleneck, we introduce an LLM-based semantic augmenter that enriches raw inputs by generating language-based abstractions that capture deeper contextual meaning. These abstractions expose latent task-relevant features, enabling robust memory indexing. In this section, we describe how these representations are generated, stored, and queried to support more efficient and generalizable episodic memory.

Representation Generation via LLMs. We construct semantic state representations by integrating commonsense knowledge through an LLM-based semantic augmenter as LLM_{aug}. Specifically, given a raw state s, the augmenter takes a structured prompt that includes the environment description, the raw observation, and the task instruction. It outputs a language-form abstraction.

$$\mathbf{l}(s) = LLM_{\text{aug}}(s), \tag{4}$$

This module is implemented via prompting and does not require additional fine-tuning. A complete template is given in the Appendix E. To enable efficient retrieval, we further compute a dense state embedding $\mathbf{z}(s_t)$ via a fixed pretrained text encoder f_{enc} (Reimers and Gurevych, 2019)

$$\mathbf{z}(s_t) = f_{\text{enc}}(\mathbf{l}(s)), \tag{5}$$

which encapsulates the textual representation of the state. These representations are more semantically meaningful and provide more information, making them ideal for indexing and retrieval in episodic memory.

Representation Storage and Memory Update. To implement episodic memory control, we maintain an episodic memory $\mathcal{M}$ that stores representations in triplets, each triplet consisting of a state embedding, the action that has historically achieved the highest observed return, and its return:

$$\mathcal{M} = \big(\mathbf{z}(s),\, a^*,\, \hat{Q}(s, a^*)\big), \tag{6}$$

$$a^* = \arg\max_a \hat{Q}(s, a), \tag{7}$$

where a^* is the action with the highest recorded return from s, and $\hat{Q}(s, a^*)$ is the maximum value.

At each timestep t, the agent observes the current state s_t and stores the corresponding state embedding $\mathbf{z}(s_t)$, the taken action a, and the received reward r in an episodic data buffer. Only after the episode ends is $\hat{Q}$ computed based on the collected data, and the episodic memory is then updated accordingly. We insert new entries for unseen states, and update existing ones only if the current return exceeds the stored value.

Memory Retrieval. We perform a KNN search in the dense state embedding space with $k = 1$. The LLM-based semantic augmenter enriches the representation with contextual meaning and latent information beyond the surface-level observation, enabling the embedding space to cluster *functionally similar* states. As a result, retrieving a single nearest neighbor is often sufficient and helps avoid irrelevant or misleading matches. This clustering effect arises from expressing goal-relevant relations and affordances in natural language, which are then encoded into dense vector representations. Let $\mathbf{z}_t$ be the query embedding at time t and $\mathbf{z}_i$ be the embedding stored with the i-th memory entry $(\mathbf{z}_i, a_i^*, \hat{Q}(s_i, a_i^*)) \in \mathcal{M}$. We measure similarity with cosine similarity:

$$\text{sim}(\mathbf{z}_t, \mathbf{z}_i) = \frac{\mathbf{z}_t^\top \mathbf{z}_i}{\|\mathbf{z}_t\|_2 \, \|\mathbf{z}_i\|_2}. \tag{8}$$

The k-nearest neighbors of s are then

$$\mathcal{N}_k(s) = \text{TopK}_{i \in \mathcal{M}} \ \text{sim}(\mathbf{z}_t, \mathbf{z}_i), \tag{9}$$

and in our experiments we set $k = 1$. The retrieved memory tuple is $(\mathbf{z}_i, a_i^*, \hat{Q}(s_i, a_i^*))$.

2.2 How to Address the Retrieval Dilemma?

To resolve the retrieval dilemma, we introduce a two-part mechanism: (i) a World-Graph working memory that maintains structured summaries for short-term reasoning, and (ii) a critical-state recognizer that selectively triggers memory recall at decision-critical moments. Together, these modules transform retrieval from passive matching into active, context-aware decisions.

World-Graph Working Memory. Prior work has shown that combining working memory with episodic memory can improve generalization in reinforcement learning agents (Fortunato et al., 2019). Building on this insight, we introduce a World-Graph working memory to maintain a structured representation of the agent's interaction history. The environment is represented as a dynamic graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}), \tag{10}$$

where each node $v \in \mathcal{V}$ corresponds to a distinct and persistent environment configuration. Edges $e \in \mathcal{E}$ represent actions that induce transitions between configurations, and $\mathcal{X}$ stores configuration-specific attributes. This design allows the graph to compactly encode the agent's accumulated knowledge of environment dynamics.

At each timestep t, the World-Graph module updates $\mathcal{G}_t$ based on the current state s_t and past interactions, adding a new node only when a previously unseen environment configuration is encountered. Edges are created or updated to reflect action-induced transitions, while node attributes are refined as new information becomes available. For example, in navigation tasks, a new node is created only when the agent enters a previously unexplored room, whereas movements within the same room update the current node. Through this incremental construction process, the World-Graph captures the agent's evolving understanding of environment dynamics and provides structured support for reasoning and decision-making.

Critical-State Recognizer. While the World-Graph working memory offers a real-time spatial abstraction of the current environment, and the

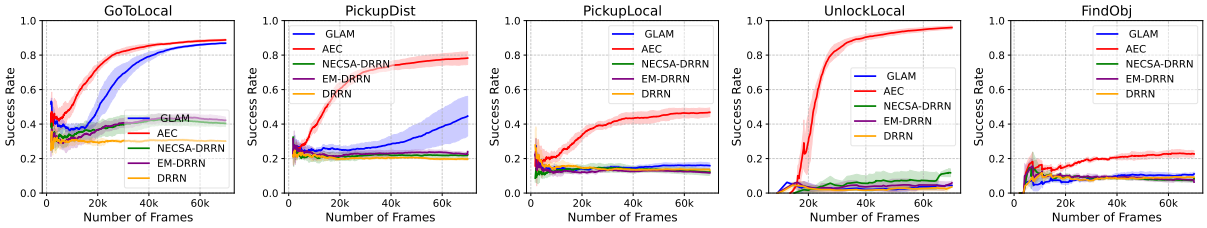

Figure 2: Learning curves (success rate vs. training frames) for five BabyAI-Text tasks comparing AEC with four baseline methods. For each method the solid line shows the mean success rate over three seeds, and the surrounding shaded region marks one standard deviation above and below that mean. AEC consistently demonstrates faster convergence and higher final performance across tasks.

episodic memory preserves valuable past experiences, the retrieval dilemma remains unresolved: when should the agent consult memory, and when should it rely on its current understanding? To address this, we introduce a critical-state recognizer, powered by a large language model, which identifies moments when episodic recall is likely to influence future outcomes. By detecting such decision-critical states, the agent can selectively trigger memory retrieval only when it is beneficial, thereby improving efficiency and avoiding unnecessary or distracting recall. This mechanism ensures that historical experience is injected precisely at important decision points, transforms memory access from passive similarity-based retrieval into targeted, decision-aware recall.

Definition 1. (Critical State) A state s_t is critical if the choice of action at s_t can substantially change the remaining cumulative return. Formally,

$$\Delta(s_t) = \max_a Q(s_t, a) - \min_a Q(s_t, a). \quad (11)$$

A state is typically classified as critical when $\Delta(s_t) > \tau_Q$, where τ_Q is a preset hyperparameter.

Computing $\Delta(s_t)$ requires a trained Q-network and a tuned threshold τ_Q, which is brittle and task-specific. Instead, we use a large language model to directly decide criticality from the semantic abstraction $LLM_{\text{aug}}(s_t)$, enabling generalization without retraining or hyperparameter tuning:

$$c(s_t) = LLM_{\text{crit}}(LLM_{\text{aug}}(s_t)) \in \{0, 1\}. \quad (12)$$

Exploitation. If the current state is classified as critical, the agent triggers an episodic memory retrieval. The agent queries the episodic memory with $f_{\text{enc}}(LLM_{\text{aug}}(s_t))$. As described in Section 2.1, we perform an exact KNN search ($k = 1$) over $\mathcal{M}$. If a matching entry exists, the agent executes the stored action; otherwise, it samples an action at random:

$$a_t = \begin{cases} a_i^*, & \text{if } (\mathbf{z}_i, a_i^*, \hat{Q}(s_i, a_i^*)) \in \mathcal{M}, \\ \text{Uniform}(\mathcal{A}), & \text{otherwise}, \end{cases} \quad (13)$$

where $\mathcal{A}$ denotes the action space.

Exploration. If the state is not critical, the agent proceeds with exploration. Given the current state s_t and the up-to-date World-Graph working memory $\mathcal{G}_t$, it chooses the next action by reasoning over these two sources of information. In practice, this step is implemented by a decision LLM that takes $(s_t, \mathcal{G}_t)$ as input and outputs an action.

3 Experiments

We assess AEC through the following research questions:

RQ1 (Data Efficiency): Can AEC improve data efficiency?

RQ2 (Generalization): How well does AEC generalize across different tasks and environments?

RQ3 (Representation Bottleneck): Does LLM-based semantic augmenter produce better state representations for episodic memory retrieval?

RQ4 (Retrieval Dilemma): Is selective episodic retrieval more effective than indiscriminate memory access?

3.1 Experimental Setup

Environment. We evaluate Agentic Episodic Control on the BabyAI-Text benchmark (Carta et al., 2023), a textual variant of BabyAI (Chevalier-Boisvert et al., 2019) that reformulates grid-based instruction-following tasks into a controlled yet challenging testbed for grounded language understanding, semantic decision-making, and generalization in reinforcement learning.

Our evaluation covers five BabyAI-Text environments that probe complementary capabilities: *Go-*

Setting	Method	GoToLocal	PickupDist	PickupLocal	UnlockLocal	FindObj
No Change	DRRN	0.13 ± 0.02	0.14 ± 0.02	0.01 ± 0.02	/	0.03 ± 0.06
	EM-DRRN	0.24 ± 0.09	0.15 ± 0.02	0.04 ± 0.01	/	0.03 ± 0.06
	NECSA-DRRN	0.18 ± 0.02	0.05 ± 0.03	0.01 ± 0.01	/	/
	GLAM	**0.91 ± 0.08**	0.69 ± 0.03	0.18 ± 0.04	0.01 ± 0.01	0.13 ± 0.06
	AEC	0.84 ± 0.02	**0.75 ± 0.17**	**0.45 ± 0.02**	**0.95 ± 0.04**	**0.23 ± 0.02**
New Object	DRRN	0.08 ± 0.02↓0.05	0.11 ± 0.03↓0.03	/↓0.01	/ •	/↓0.03
	EM-DRRN	0.10 ± 0.07↓0.14	0.09 ± 0.09↓0.06	/↓0.04	/ •	/↓0.03
	NECSA-DRRN	0.19 ± 0.03↑0.01	0.02 ± 0.02↓0.03	0.02 ± 0.00↑0.01	/ •	/ •
	GLAM	**0.85 ± 0.10**↓0.06	0.61 ± 0.04↓0.08	0.16 ± 0.06↓0.02	0.01 ± 0.01•	0.12 ± 0.03↓0.01
	AEC	0.83 ± 0.02↓0.01	**0.71 ± 0.12**↓0.04	**0.52 ± 0.06**↑0.07	**0.95 ± 0.02**•	**0.20 ± 0.03**↓0.03

Table 1: Final success rates (mean ± std over 3 seeds) after 70K environment frames on five BabyAI-Text environments, evaluated under two settings: **No Change** (seen objects) and **New Object** (unseen objects). "/" indicates the method failed to solve the task. Performance differences between the two settings are indicated as follows: ↑ (increase), ↓ (decrease), and • (no change).

ToLocal, *PickupDist*, *PickupLocal*, *UnlockLocal*, and *FindObj*.Following the protocol in (Carta et al., 2023), we also test the agent's ability to generalize to unseen instructions and object types. We refer to the standard setting as *No change* and the version containing novel, unseen target objects as *New object*. More details are given in Appendix C. AEC is built on top of the `Qwen2.5-32B-Instruct` (Team, 2024) model, a powerful large language model with strong instruction-following capabilities.

Baselines. We compare AEC against a range of baselines, covering both reinforcement learning and large language model–augmented approaches. **DRRN** (He et al., 2016) serves as a standard RL baseline with competitive performance in interactive text environments (Wang et al., 2022), including the BabyAI-Text benchmark (Carta et al., 2023). **EM-DRRN** extends DRRN with an episodic memory module, following the design of EMDQN (Lin et al., 2018), to improve sample efficiency. **NECSA-DRRN** incorporates NECSA (Li et al., 2023) into DRRN, discretizing the state-action space into a grid-like structure for efficient memory-based value estimation. **GLAM** (Carta et al., 2023) fine-tunes a large language model with PPO to act as a policy directly in textual environments, serving as a strong LLM-based baseline.

3.2 RQ1: Data Efficiency

We benchmarked AEC against four baselines on five BabyAI-Text environments. As shown in Figure 2, AEC consistently achieves higher success rates with significantly fewer interactions. In the obstacle navigation environment GoToLocal, AEC surpasses an 80% success rate within 25K frames, while GLAM requires approximately 50K frames to reach a comparable level (75%), demonstrating nearly a 2× improvement in sample efficiency. Although EM-DRRN and NECSA-DRRN benefit from episodic memory to slightly accelerate learning, they still fall far behind AEC. In the object pickup environment PickupDist, for example, GLAM requires 60K frames to reach a 40% success rate, whereas AEC achieves this performance within just 10K frames—representing a 6× gain in data efficiency. For the multi-step reasoning environment UnlockLocal, AEC is the only method capable of solving the task, surpassing a 90% success rate within 40K frames. All baselines fail to exceed 15% even after 70K frames, highlighting AEC's strong generalization and planning capabilities. In the most challenging environment, FindObj, which involves multi-room exploration under sparse rewards, AEC again stands out as the only method to exceed a 20% success rate within 70K frames.

Taken together, these results clearly demonstrate that AEC significantly improves data efficiency across a diverse set of language-conditioned reasoning and navigation tasks.

3.3 RQ2: Generalization

Cross-Task. We evaluate AEC's generalization ability across different tasks settings by comparing its performance under two evaluation conditions: No Change, where the agent encounters familiar objects seen during training, and New Object, where novel target objects are introduced at test time.

As shown in Table 1, AEC demonstrates strong generalization across all five BabyAI-Text tasks. While it performs slightly below GLAM on GoToLocal, it consistently outperforms all baselines

on the remaining tasks under both evaluation settings. Notably, AEC is the only method that maintains high performance even when evaluated with previously unseen objects. Despite the distribution shift introduced in the New Object setting, AEC retains most of its performance. For example, on PickupLocal, AEC even exceeds its performance under the No Change condition, achieving a relative improvement of +16% (from 0.45 to 0.52). In more challenging scenarios like FindObj, AEC still retains 87% of its original performance (dropping only slightly from 0.23 to 0.20), whereas other baselines degrade significantly or fail entirely. These results confirm that AEC's integration of language-grounded abstraction and structured memory enables it to generalize effectively to novel task configurations with minimal performance loss.

Cross-Environment. We investigate whether AEC's episodic memory enables cross-environment generalization. To this end, we conducted a transfer experiment in which the agent attempted to solve a task using episodic memories retrieved from a different environment. Specifically, we evaluated performance on the GotoLocal and PickupLocal environments under three memory settings: raw states from the same environment, memory from the same environment, and memory from a different environment.

As shown in Table 2, reusing memory from a related environment leads to notable improvements over using raw states. For example, using PickupLocal-derived memory to solve GotoLocal yields a success rate of 0.75, substantially higher than the 0.49 obtained with raw states. Similarly, applying GotoLocal memory to PickupLocal improves performance from 0.30 to 0.39. Interestingly, we observe an asymmetry in transferability: experiences learned from the more complex PickupLocal environment benefit GotoLocal more than the reverse. This suggests that solving more demanding environments may yield richer, more generalizable memory representations, providing stronger priors when applied to simpler environments. These results indicate the potential for cross-environment memory reuse in AEC and open avenues for future work on compositional generalization.

3.4 RQ3: Representation Bottleneck

To evaluate whether our method alleviates the representation bottleneck in episodic reinforcement learning, we perform a controlled representation analysis on states with known decision semantics. We construct a balanced set of 300 states by randomly sampling 100 states for each of three manually identified optimal actions, enabling a direct test of whether similarity in the embedding space aligns with *decision equivalence*.

We compare four state representations: SentenceTransformer embeddings of LLM-augmented semantic representations, SentenceTransformer embeddings of raw textual states, embeddings from a DRRN encoder trained for 70K environment frames, and embeddings from an untrained DRRN encoder. For each representation, we perform KNN retrieval and measure *KNN retrieval accuracy*, defined as the proportion of retrieved neighbors that share the same optimal action.

As shown in Figure 3, LLM-based representations achieve consistently high retrieval accuracy across all values of k, exceeding 95% at $k = 1$ and remaining above 92% at $k = 10$. DRRN representations trained for 70K frames show only a modest improvement over untrained DRRN embeddings. We note that with substantially more training data and interaction, DRRN-style encoders may learn stronger task-specific representations. However, under the low-data regime considered here, shallow and partially trained encoders fail

Memory Source	Target Task	
	GotoLocal	PickupLocal
Raw States	0.49	0.30
GotoLocal Memory	0.83	0.39
PickupLocal Memory	0.75	0.52

Table 2: Cross-task memory transfer results. Success rates on GotoLocal and PickupLocal tasks when using episodic memory encoded from different tasks. "Raw States" denotes raw states without LLM-based semantic augmenter from the same task.

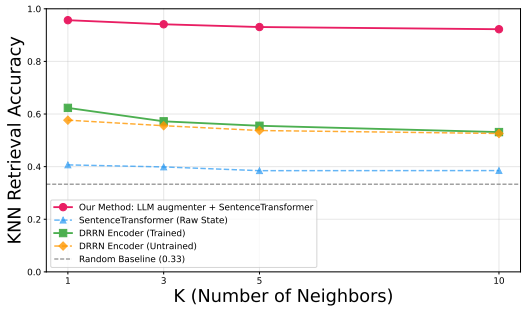

Figure 3: KNN retrieval accuracy for different state representations.

LLM	GoToLocal	PickupLocal	FindObj
Qwen2.5-7B-Instruct	0.79 ± 0.03	0.43 ± 0.02	0.13 ± 0.01
Qwen2.5-32B-Instruct	0.84 ± 0.02	0.45 ± 0.02	0.23 ± 0.02

Table 3: Robustness across different LLM backbones on the three BabyAI-Text tasks (100 test environments). Values are means over three seeds; ± denotes one standard deviation.

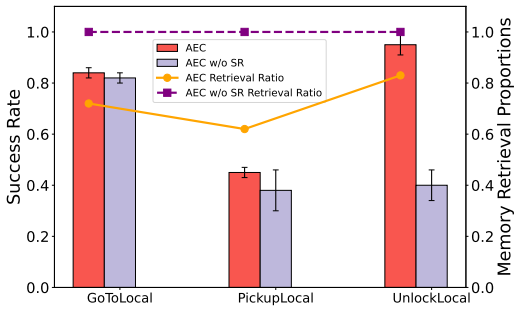

Figure 4: Comparison of success rates and memory retrieval proportions on three BabyAI-Text environments. SR denotes Selective Retrieval. Error bars represent one standard deviation.

to align embedding similarity with decision relevance, whereas LLM-based semantic augmenter yields highly action-consistent neighborhoods.

3.5 RQ4: Retrieval Dilemma

To investigate the impact of the retrieval dilemma on decision quality, we compare AEC against a variant that queries episodic memory indiscriminately at every timestep (AEC w/o selective retrieval). Crucially, both agents utilize the identical episodic memory content to ensure a fair comparison. Figure 4 reports task success rates and memory retrieval proportions across three environments: GoToLocal, PickupLocal, and UnlockLocal.

AEC consistently outperforms the indiscriminate baseline across all tasks. This performance gap is particularly pronounced in UnlockLocal, where the baseline's success rate drops significantly. These results indicate that accessing episodic memory at every step is detrimental to decision accuracy, potentially due to the interference caused by retrieving information at non-critical states. Consequently, selectively invoking episodic memory is essential for maximizing task success and robustness in complex environments.

3.6 Ablation studies

We conduct ablation studies on all five BabyAI-Text environments to assess the contribution of

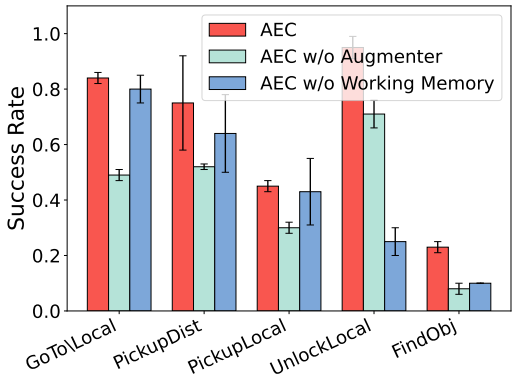

Figure 5: Ablation study results across five BabyAI-Text environments. Error bars represent one standard deviation.

AEC's components. As shown in Figure 5, removing either module consistently degrades performance across tasks, demonstrating their complementary roles. In particular, working memory is crucial for FindObj, which features extremely sparse rewards, where selective memory retrieval is essential to avoid uninformative recall. These results confirm that both the LLM-based semantic augmenter and the working memory are necessary for AEC's strong performance, enabling semantic generalization and long-horizon reasoning.

To evaluate robustness across backbone models, we repeat experiments using the smaller `Qwen2.5-7B-Instruct` model on three tasks of varying difficulty, with three random seeds per task. Table 3 reports mean success rates over 100 test environments. AEC achieves consistent performance with both backbones, while the 32B model performs better overall, indicating that AEC further benefits from the richer knowledge of larger language models.

4 Conclusions

We introduced Agentic Episodic Control, a framework that integrates LLM-based semantic abstraction with structured memory to resolve the representation bottleneck and retrieval dilemma in episodic RL. By replacing passive retrieval with decision-aware recall, AEC achieves significant improvements in sample efficiency and generalization, consistently outperforming strong baselines in complex, long-horizon tasks. These results underscore the effectiveness of combining semantic understanding with active memory mechanisms, offering a robust path toward more data-efficient autonomous agents.

Limitations

Despite its strong performance, AEC incurs higher inference latency compared to traditional RL baselines due to the computational cost of LLM queries and structured memory updates. This overhead may limit its deployment in real-time, high-frequency control settings. Future work will focus on mitigating these costs through techniques such as knowledge distillation into smaller models or caching intermediate semantic representations.

References

Vinamra Benara, Chandan Singh, John X. Morris, Richard J. Antonello, Ion Stoica, Alexander Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking llms questions. In *Advances in Neural Information Processing Systems*.

Siddhant Bhambri, Amrita Bhattacharjee, Subbarao Kambhampati, and 1 others. 2024. Efficient reinforcement learning via large language model-based search. In *NeurIPS 2024 Workshop on Open-World Agents*.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. 2016. Model-free episodic control. *arXiv preprint arXiv:1606.04460*.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*.

Kangkang Chen, Zhongxue Gan, Siyang Leng, and Chun Guan. 2022. Deep reinforcement learning with parametric episodic memory. In *2022 International Joint Conference on Neural Networks*.

Siwei Chen, Anxing Xiao, and David Hsu. 2023. Llm-state: Open world state representation for long-horizon task planning with large language model. *arXiv preprint arXiv:2311.17406*.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. Babyai: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*.

Longchao Da, Minquan Gao, Hao Mei, and Hua Wei. 2024. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*.

Howard Eichenbaum. 2004. Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, 44(1):109–120.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, and 1 others. 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53.

Meire Fortunato, Melissa Tan, Ryan Faulkner, Steven Hansen, Adrià Puigdomènech Badia, Gavin Buttimore, Charles Deck, Joel Z Leibo, and Charles Blundell. 2019. Generalization of reinforcement learners with working and episodic memory. In *Advances in Neural Information Processing Systems*.

Steven Hansen, Alexander Pritzel, Pablo Sprechmann, André Barreto, and Charles Blundell. 2018. Fast deep reinforcement learning using online adjustments from the past. In *Advances in Neural Information Processing Systems*.

Ji He, Mari Ostendorf, Xiaodong He, Jianshu Chen, Jianfeng Gao, Lihong Li, and Li Deng. 2016. Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Ezgi Korkmaz. 2024. A survey analyzing generalization in deep reinforcement learning. *arXiv preprint arXiv:2401.02349*.

B. A. Kuhl and A. D. Wagner. 2009. *Strategic Control of Memory*, pages 437–444. Elsevier Ltd.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.

Su Young Lee, Sung-Ik Choi, and Sae-Young Chung. 2019. Sample-efficient deep reinforcement learning via episodic backward update. In *Advances in Neural Information Processing Systems*.

Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. 2024. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, and 1 others. 2022. Pre-trained language models for interactive decision-making. In *Advances in Neural Information Processing Systems*.

Zhuo Li, Derui Zhu, Yujing Hu, Xiaofei Xie, Lei Ma, Yan Zheng, Yan Song, Yingfeng Chen, and Jianjun Zhao. 2023. Neural episodic control with state abstraction. In *International Conference on Learning Representations*.

Dayang Liang, Yaru Zhang, and Yunlong Liu. 2024. Episodic reinforcement learning with expanded state-reward space. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

Zichuan Lin, Tianqi Zhao, Guangwen Yang, and Lintao Zhang. 2018. Episodic memory deep q-networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.

Daniel J Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, and 1 others. 2023. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263.

Hyungho Na, Yunkyeong Seo, and Il-chul Moon. 2024. Efficient episodic memory utilization of cooperative multi-agent reinforcement learning. In *International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Shaohui Peng, Xing Hu, Qi Yi, Rui Zhang, Jiaming Guo, Di Huang, Zikang Tian, Ruizhi Chen, Zidong Du, Qi Guo, and 1 others. 2024. Hypothesis, verification, and induction: grounding large language models with self-driven skill learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. 2017. Neural episodic control. In *International conference on machine learning*.

Michelle M. Ramey, John M. Henderson, and Andrew P. Yonelinas. 2022. Episodic memory processes modulate how schema knowledge is used in spatial memory decisions. *Cognition*, 225:105111.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. 2021. Pretraining representations for data-efficient reinforcement learning. In *Advances in Neural Information Processing Systems*.

Chak Lam Shek, Xiyang Wu, Wesley A. Suttle, Carl E. Busart, Erin G. Zaroukian, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. 2024. LANCAR: leveraging language for context-aware robot locomotion in unstructured environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Ruizhe Shi, Yuyao Liu, Yanjie Ze, Simon Shaolei Du, and Huazhe Xu. 2024. Unleashing the power of pre-trained language models for offline reinforcement learning. In *International Conference on Learning Representations*.

Yash Shukla, Wenchang Gao, Vasanth Sarathy, Alvaro Velasquez, Robert Wright, and Jivko Sinapov. 2024. Lgts: Dynamic task sampling using llm-generated sub-goals for reinforcement learning agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Weihao Tan, Wentao Zhang, Shanqi Liu, Longtao Zheng, Xinrun Wang, and Bo An. 2024. Twosome: An efficient online framework to align llms with embodied environments via reinforcement learning. *International Journal of Artificial Intelligence and Robotics Research*, 01(02):2450004.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Luong Quoc Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, and 1 others. 2021. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596.

Boyuan Wang, Yun Qu, Yuhang Jiang, Jianzhun Shao, Chang Liu, Wenming Yang, and Xiangyang Ji. 2024.

Llm-empowered state representation for reinforcement learning. In *International Conference on Machine Learning*.

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

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *arXiv preprint arXiv:2312.17661*.

Yue Wu, Yewen Fan, Paul Pu Liang, Amos Azaria, Yuanzhi Li, and Tom Mitchell. 2023. Read and reap the rewards: learning to play atari with the help of instruction manuals. In *Advances in Neural Information Processing Systems*.

Gui Xue. 2018. The neural representations underlying human episodic memory. *Trends in Cognitive Sciences*, 22(6):544–561.

Xue Yan, Yan Song, Xinyu Cui, Filippos Christianos, Haifeng Zhang, David Henry Mguni, and Jun Wang. 2023. Ask more, know better: Reinforce-learned prompt questions for decision making with large language models. *arXiv preprint arXiv:2310.18127*.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building cooperative embodied agents modularly with large language models. In *International Conference on Learning Representations*.

Guangxiang Zhu, Zichuan Lin, Guangwen Yang, and Chongjie Zhang. 2020. Episodic reinforcement learning with associative memory. In *International Conference on Learning Representations*.

A Ethical Considerations

This study investigates the role of large language models as components within reinforcement learning agents, and all experimental results are obtained by evaluating agent performance in controlled environments. In addition, an LLM was used as an auxiliary tool during manuscript preparation, including improving grammar and clarity of the text. The LLM was not used to generate experimental data, design algorithms, analyze results, or draw scientific conclusions. All research ideas, methods, experiments, analyses, and conclusions were conceived and conducted by the authors.

B Related Works

B.1 Neural Episodic Control

Episodic control refers to capturing and retaining the most influential experiences during the learning process to guide an agent's decision-making. Model-free episodic control (MFEC) (Blundell et al., 2016) employs episodic memory buffers that hold experience tuples, compressing states via Gaussian random projection or variational autoencoders. NEC (Pritzel et al., 2017) introduced a differentiable neural dictionary that stores past experiences, with convolutional networks used for state embedding. Subsequent studies modify parameterised models through episodic memory (Lin et al., 2018; Hansen et al., 2018; Chen et al., 2022; Liang et al., 2024; Lee et al., 2019). Recently, research has shifted to state representation in neural episodic control. Episodic Reinforcement Learning with Associative Memory (ERLAM) (Zhu et al., 2020) associates related trajectories to form policies with efficient reasoning and propagates new values through a graph built over memory items. Neural Episodic Control with State Abstraction (NECSA) (Li et al., 2023) performs state abstraction on grid-based environments to exploit topological structures and enhance policy learning and generalization. Efficient Episodic Memory Utilization (EMU) (Na et al., 2024) constructs semantically coherent memory embeddings with a trainable encoder–decoder. Cooperative Embodied Language Agent (CoELA) (Zhang et al., 2024) maintains three memory systems—semantic, episodic, and procedural—that are updated primarily by logging raw action and dialogue histories. However, human episodic memory encodes experiences with prior knowledge and semantic detail (Xue, 2018; Ramey et al., 2022). We therefore introduce an LLM-based semantic augmenter, aligned with this human characteristic to enhance decision-making.

B.2 LLMs for Reinforcement Learning

Effectively enhancing reinforcement learning with large language models is a promising research area. Several approaches have been explored to leverage LLMs in this field. One method is based on a hierarchical framework, in which LLMs are used to decompose complex tasks and generate high-level plans, which are then executed by low-level controllers (Peng et al., 2024; Shukla et al., 2024). Another approach utilizes LLMs to design reward functions, simplifying the typically labor-intensive process of reward function formulation (Kwon et al., 2023; Wu et al., 2023; Li et al., 2024; Bhambri et al., 2024). Some studies directly use LLMs as behavioral policies, training them via RL to interact with the environment (Yan et al., 2023; Shi et al., 2024; Li et al., 2022; Carta et al., 2023; Tan et al., 2024). For example, Grounded Language Models (GALM) (Carta et al., 2023) and Twosome (Tan et al., 2024) apply LLMs to text-based games. Our work differs from these approaches by leveraging the semantic understanding of LLMs to enhance episodic memory embedding, while dynamically coordinating with graph-structured working memory for real-time reasoning. This tight integration enables more human-like generalization and sample efficiency compared to existing LLM-RL hybrids.

B.3 LLM-based State Embedding

Recent advances demonstrate the versatile capabilities of large language models in enriching state representations across diverse applications (Da et al., 2024; Shek et al., 2024; Chen et al., 2023; Wang et al., 2024; Benara et al., 2024). Language for context-aware robot locomotion (LANCAR) (Shek et al., 2024) employs an LLM-based contextual information translation module to interpret high-level information from the environment into contextual embeddings accessible to RL agents. LLM-State (Chen et al., 2023) leveraged the inherent contextual understanding and historical action reasoning capabilities of LLMs to provide continuous expansion and updates of object properties. Question-answering embeddings (QA-Emb) (Benara et al., 2024) is a method that leverages LLMs to answer a series of yes/no questions for generating interpretable state embeddings. Our method integrates a LLM-based semantic encoder directly

into the agent"s core memory architecture, enabling deep semantic grounding of both episodic memory storage and retrieval processes.

C Environment Details

We evaluate the proposed Agentic Episodic Control architecture on the BabyAI-Text benchmark (Carta et al., 2023). The wrapper (Carta et al., 2023) built on top of BabyAI removes reliance on visual perception and allows us to isolate the benefits of language-driven episodic memory and structured reasoning. In this environment, the agent is provided with a natural language instruction and a textual description of its local egocentric observation. This observation represents a forward-facing view in text form, covering up to seven adjacent tiles directly in front of the agent. This design simulates partial observability and tests the agent's ability to reason over limited, language-grounded sensory input.

Our evaluation covers five **BabyAI-Text** environments that probe complementary capabilities:

- *GoToLocal*: The agent must navigate to a specified object in a room with distractors, assessing spatial grounding and distractor resistance.
- *PickupDist*: Requires picking up a specified object (by type and/or color) amid distractors, stressing fine-grained object disambiguation.
- *PickupLocal*: Builds on *PickupDist*, but describes the target via its location, testing grounding of locative language in a single-room setting.
- *UnlockLocal*: Involves fetching the correct key and unlocking a door in the current room, emphasizing multi-step reasoning and tool use.
- *FindObj*: Evaluates exploration, requiring the agent to explore across six rooms to locate a target object.

Following the protocol in (Carta et al., 2023), we also test the agent's ability to generalize to unseen instructions and object types. We refer to the standard evaluation setting as *No change*, where the object names and colors present in the environment during testing are consistent with those seen during training. However, the target instruction and room layout vary across episodes. In contrast, the *New object* setting introduces novel, previously unseen object names and colors at test time, while still varying the target and room layout in each episode.

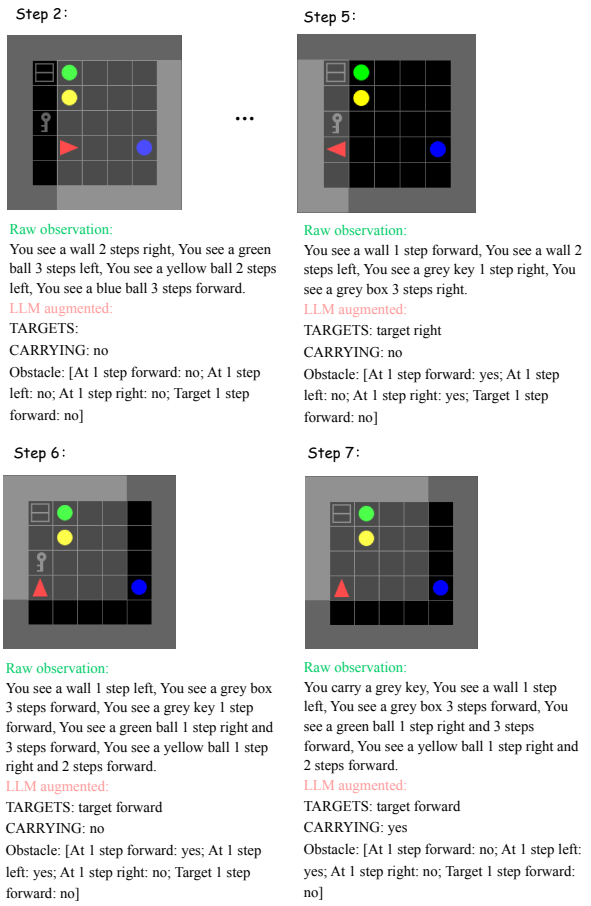

Figure 6: Case study of the LLM-based semantic augmenter in the PickupDist task (*pick up the gray box*). For selected timesteps, we show the textual description, and the LLM-augmented semantic representations.

D Case Studies

Analysis of LLM-based semantic augmenter. To illustrate the role of the LLM-based semantic augmenter, we analyze a representative episode from the PickupDist task, focusing on how raw observations are transformed into structured, decision-relevant representations.

As shown in Figure 6, raw observations consist of unstructured, surface-level descriptions that enumerate visible objects and their relative positions. While informative, these observations entangle task-relevant signals with irrelevant details, making it difficult for similarity-based retrieval to identify functionally equivalent states.

In contrast, the LLM-based semantic augmenter distills each observation into a compact representa-

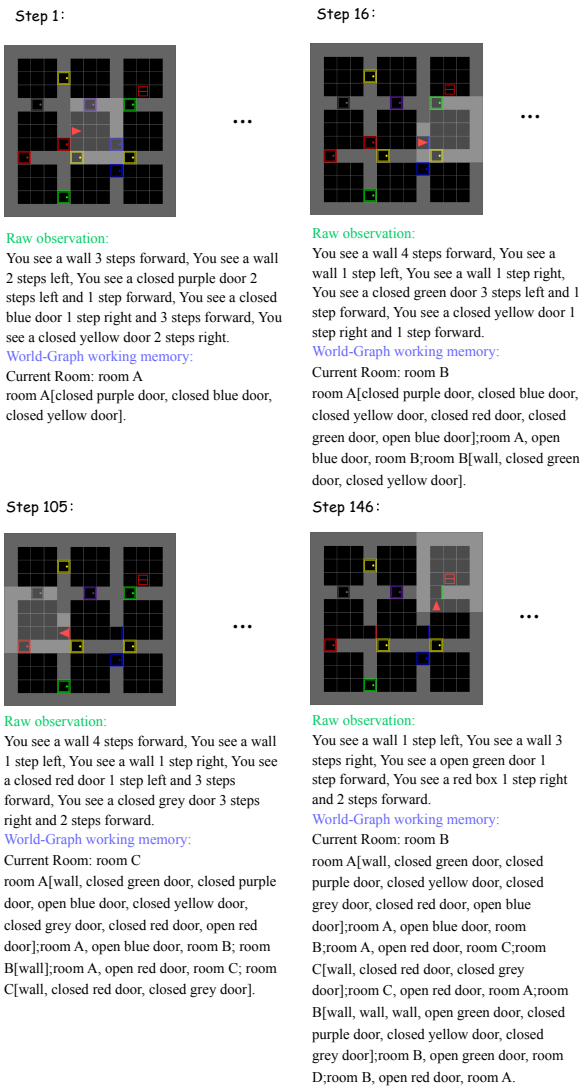

Figure 7: Selected key steps from the "pick up the box" task. Each frame illustrates the agent's textual description, and the corresponding state of its world-graph working memory. Since the environment does not provide explicit room names, the agent assigns room labels (e.g., room A, room B) based on its own exploration order and infers connectivity between rooms using door colors and relative positions.

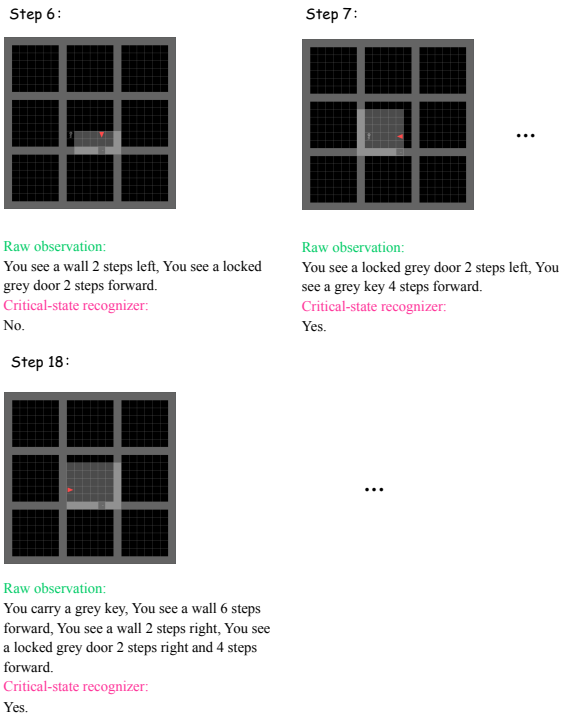

Figure 8: Performance of the critical-state recognizer in the UnlockLocal environment.

tion that explicitly highlights task-critical factors, including the relative direction of the target , the agent's carrying status, and local obstacle availability. For example, at Step 5, the augmenter correctly identifies the target as being to the right and marks the immediate right position as blocked by an obstacle, while suppressing unrelated objects such as balls that do not affect decision-making. As the episode progresses (Steps 6–7), the semantic representation remains stable despite changes in raw visual content, while dynamically updating key attributes such as the agent's carrying state after picking up the key.

This abstraction process effectively maps perceptually different but functionally similar states to nearby representations, aligning the embedding space with decision relevance rather than surface appearance. As a result, episodic memory retrieval operates over semantically meaningful keys, enabling reliable reuse of past experience across variations in object layout and observation wording. This case study demonstrates that the LLM-based semantic augmenter plays a crucial role in alleviating the representation bottleneck by transforming raw observations into task-aligned, decision-centric state representations.

Analysis of World-Graph Working Memory. To evaluate the correctness of the world-graph working memory constructed by the agent during task execution, we analyzed a representative episode from the FindObj task, specifically the "pick up the box" scenario. As illustrated in the Figure 7, we highlight several key steps in the episode, showing the visual observations, textual descriptions, and the corresponding state of the agent's world-graph.

Our analysis shows that the agent is capable of accurately building and maintaining connectivity information between rooms during exploration. For instance, by step 146, the world-graph correctly

reflects that Room A is connected to Room B via a blue door, to Room C via a red door, and that Room B is connected to Room D through a green door. This structured spatial representation allows the agent to plan paths and navigate toward the goal room effectively.

It is important to note that since the environment does not provide explicit room identifiers, the agent assigns room names based on the order of exploration and relies on room attributes to recognize its current location. This approach poses challenges, especially in dynamic environments where, for instance, door states may change—sometimes leading to attribute misidentification and local inconsistencies in the world-graph. Nevertheless, despite these uncertainties, the agent is still able to infer room connectivity accurately and reliably identify paths to the target, demonstrating robustness in spatial memory construction and planning.

Analysis of Critical-State Recognizer. The Figure 8, based on the UnlockLocal environment, illustrates the effectiveness of the critical-state recognizer. In the "open the door" task, before the agent obtains the key, the recognizer identifies observations containing the key as critical states. After acquiring the key, it shifts to recognizing locked doors as critical. This dynamic adaptation shows that the recognizer can adjust its definition of criticality in real time according to the agent's task progress. As a result, it more effectively guides the agent's balance between exploration and exploitation, transforming retrieval from passive matching into active, context-aware decision-making—thereby addressing the retrieval dilemma.

E Prompt Engineering

We introduce the four prompt templates employed in our experiments. These prompts are carefully crafted to provide structured and consistent instructions that guide the large language model's behavior within the text-based maze environment. Their design aims to enhance clarity, ensure uniform interpretation of observations, support coherent memory management, and promote rational decision-making. The complete content of each prompt is presented below.

The prompt for LLM-based semantic augmenter

```
You are MazeNavigator-GPT, an expert at reading
text-based maze observations.
  GAME DESCRIPTION
  {Game description}
  MISSION
  {Mission}
  INPUT
  {Observation}
  TASK (Do all in order, think step-by-step)
  STEP 0 - PARSE MISSION
  Analysis the game description and the
observation, determine your current target object.
If the prerequisites for certain goals are not met,
they should not be set as current target object.
Extract the exact color + item that defines the
current target object. Store it as one pair (e.g.
"red ball"). This is the current target object.
  STEP 1 - BUILD OUTPUT LIST
  Try to find the current target object in the
observation. Scan the single observation for any
mention of that current target object with the
same color. Ignore objects whose does not match
exactly. For each valid sentence: remove all
numbers (distances, counts, step sizes). Keep the
pure direction phrase (e.g. "left and forward",
"right"). Create an entry in the form: "target
<direction phrase >". Example → "target left and
forward". Preserve the order they appear in the
observation.
  STEP 2 - CARRYING ITEM?
  Output "yes" if the observation states the agent
is carrying anything; else "no".
  STEP 3 - OBSTACLE ANALYSIS
  Is there any objects or walls are there in the
three directions of the agent: "1 step forward
without left and right", "1 step left without
forward", "1 step right without forward"? If no
information provided, output "no".
  STEP 4 - TARGET ANALYSIS
  Target 1 step forward? Output "Target 1 step
forward: yes" if the current target object only at
1 step forward; else if the current target object
is at 1 step left/right and 1 step forward, output
"Target 1 step forward: no".
  OUTPUT FORMAT
  {Format requirements}
```

The prompt for World-Graph working memory construction

```
OBJECTIVE
  The main goal is to meticulously gather
information from maze exploration observations and
update a structured knowledge graph. Don"t include
any speculation. Only incorporate new, directly
observed information.
  Always retain all items and triplets from the
previous knowledge graph. Do not remove or modify
any previously recorded objects or relationships,
even if they are not observed in the current
observation.
  PREVIOUS WORLD MODEL
  {Previous world model}
  PREVIOUS OBSERVATION
  {Previous observation}
  ACTION
  {Action}
  NEW OBSERVATION
  {New observation}
  TASK
  Building and Updating the Knowledge Graph -
Follow these steps:
  STEP 0 - Room Labeling and Node Properties
```

Assign room labels sequentially as rooms are confirmed. Name the first room "room A", and label subsequent new rooms as "room B", "room C", etc. If a room has been labeled previously, reuse that label consistently. For the room you have just passed, include all explicitly observed objects in its properties. IMPORTANT: first, import any objects already memorized for that room from the previous world model. Then, add only the newly observed objects (if they are not already recorded). Do not remove any objects from previous observations.- Ignore object positions or orientations. Format: room A [item 1, item 2, item 3, ...]; ...

At the beginning of your output, include a line stating the current room label. Example: "Current Room: room A".

STEP 1 - Triplet Creation

Only perform this step if the new observation explicitly mentions a door. If the new observation does not explicitly mention a door, SKIP THIS STEP entirely and do not create any triplets. When creating triplets, strictly use the format: subject, relation, object. Here the subject is the current room. The relation is the specific door (using the color and state as explicitly mentioned in the observation). The object is the new room reached by that door.

OUTPUT FORMAT

{Format requirements}

The prompt for critical-state recognizer

You are MazeNavigator-GPT, a specialist in solving navigation games.

GAME DESCRIPTION

{Game description}

MISSION

{Mission}

CURRENT OBSERVATION

{Current observation after augmentation}

TASK

Based on the current observation, determine if there is a target direction. If there is, output "yes". If not, output "no". Do not provide any additional explanations, comments, or text.

The prompt for exploration strategy using World-Graph working memory

You are MazeNavigator-GPT, a specialist in solving navigation games by selecting the optimal action based on the mission, world model, and observations. You enjoy exploring unknown environments.

GAME DESCRIPTION

{Game description}

MISSION

{Mission}

ACTION SPACE:

{Action space}

WORLD MODEL

{World model}

HISTORY

{History}

CURRENT OBSERVATION

{Current observation}

TASK

Based on the world model, history and current observation, select the single most effective action that brings you closer to the mission goal or advances environmental exploration(if no mission objective is present). Be careful to avoid falling into a simple loop of exploration solely through turning left or right;try considering moving forward, picking up items, or other actions to break the stalemate. If you notice that repeating the same actions in consecutive similar scenarios fails to make progress, prioritize trying other options.

OUTPUT FORMAT

{Format requirements}