SYNAPSE: Empowering LLM Agents with Episodic-Semantic Memory via Spreading Activation

Hanqi Jiang[1,2*], **Junhao Chen**[1,2*], **Yi Pan**[1,2], **Ling Chen**[3], **Weihang You**[1,2], **Yifan Zhou**[1,2], **Ruidong Zhang**[1,2], **Yohannes Abate**[4], **Tianming Liu**[1,2†]

[1]GyriQAI, Inc.

[2]School of Computing, University of Georgia, Athens

[3]Department of Biosystems Engineering and Soil Science, University of Tennessee, Knoxville

[4]Department of Physics and Astronomy, The University of Georgia, Athens

Abstract

While Large Language Models (LLMs) excel at generalized reasoning, standard retrieval-augmented approaches fail to address the disconnected nature of long-term agentic memory. To bridge this gap, we introduce SYNAPSE (**Syn**ergistic **A**ssociative **P**rocessing & **S**emantic **E**ncoding), a unified memory architecture that transcends static vector similarity. Drawing from cognitive science, SYNAPSE models memory as a dynamic graph where relevance emerges from spreading activation rather than pre-computed links. By integrating lateral inhibition and temporal decay, the system dynamically highlights relevant sub-graphs while filtering interference. We implement a Triple Hybrid Retrieval strategy that fuses geometric embeddings with activation-based graph traversal. Comprehensive evaluations on the LoCoMo benchmark show that SYNAPSE significantly outperforms state-of-the-art methods in complex temporal and multi-hop reasoning tasks, offering a robust solution to the "Contextual Tunneling" problem. Our code and data will be made publicly available upon acceptance.

1 Introduction

The evolution of Large Language Models (LLMs) from static responders to autonomous agents necessitates a fundamental rethinking of memory architecture (Park et al., 2023; Yao et al., 2023; Schick et al., 2023). While LLMs demonstrate remarkable reasoning within finite context windows, their agency is brittle without the ability to accumulate experiences and maintain narrative coherence over long horizons (Gutiérrez et al., 2024; Izacard et al., 2023). The predominant solution, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), externalizes history into vector databases, retrieving information based on semantic similarity (Guu et al., 2020; Asai et al., 2024). While effective for factual lookup (Borgeaud et al., 2022), standard RAG imposes a critical limitation on reasoning agents: it treats memory as a static library to be indexed, rather than a dynamic network to be reasoned over (Gutiérrez et al., 2024; Zhu et al., 2025).

We argue that existing systems suffer from **Contextual Isolation**, a failure mode stemming from the implicit **Search Assumption**: that the relevance of a past memory is strictly determined by its semantic proximity to the current query (Zhu et al., 2025; Edge et al., 2025; Sarthi et al., 2024). This assumption collapses in scenarios requiring causal or transitive reasoning. Consider a user asking, *"Why am I feeling anxious today?"*. A vector-based system might retrieve recent mentions of "anxiety," but fail to surface a schedule conflict logged weeks prior. Although this conflict is the root cause, it shares no lexical or embedding overlap with the query. While hierarchical frameworks such as MemGPT (Packer et al., 2024) improve context management, they remain bound by query-driven retrieval, unable to autonomously surface structurally related yet semantically distinct information.

To bridge this gap, we draw inspiration from cognitive science theories of Spreading Activation (Collins and Loftus, 1975; Anderson, 1983), which posit that human memory retrieval is not a search process, but a propagation of energy. Accessing one concept naturally activates semantically, temporally, or causally linked concepts without explicit prompting.

We introduce SYNAPSE, a brain-inspired architecture that reimagines agentic memory. Unlike flat vector stores, SYNAPSE constructs a Unified Episodic-Semantic Graph, where raw interaction logs (episodic nodes) are synthesized into abstract concepts (semantic nodes). Retrieval in SYNAPSE is governed by activation dynamics: input signals inject energy into the graph, which propagates

[*]Equal contribution

[†]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 22020–22036

through temporal and causal edges. This mechanism enables the system to prioritize memories that are structurally salient to the current context, such as the aforementioned schedule conflict, even when direct semantic similarity is absent. To ensure focus, we implement lateral inhibition, a biological mechanism that suppresses irrelevant distractors.

We evaluate SYNAPSE on the rigorous LoCoMo benchmark (Maharana et al., 2024), which involves long-horizon dialogues averaging 16K tokens. SYNAPSE establishes a new state-of-the-art (SOTA), significantly outperforming traditional RAG and recent agentic memory systems. Notably, our activation-based approach improves accuracy on complex multi-hop reasoning tasks by up to 23% while reducing token consumption by 95% compared to full-context methods.

In summary, our contributions are as follows:

- **Unified Episodic-Semantic Graph:** We propose a dual-layer topology that synergizes granular interaction logs with synthesized abstract concepts, addressing the structural fragmentation inherent in flat vector stores.
- **Cognitive Dynamics with Uncertainty Gating:** We introduce a retrieval mechanism governed by spreading activation and lateral inhibition to prioritize implicit relevance, coupled with a "feeling of knowing" protocol that robustly rejects hallucinations.
- **SOTA Performance & Efficiency:** SYNAPSE establishes a new state-of-the-art on the LoCoMo benchmark (+7.2 F1), improving multi-hop reasoning accuracy by 23% while reducing token consumption by 95% compared to full-context methods.

2 Related Work

2.1 Memory Allocation Capabilities

Systems such as MemGPT (Packer et al., 2024), MemoryOS (Li et al., 2025), and LangMem (LangChain Team, 2024) address context limitations by optimizing memory placement via policy-based controllers or hierarchical buffers (Lewis et al., 2020; Nafee et al., 2025; Guu et al., 2020). However, these approaches treat memory items as independent textual units, lacking the mechanisms to model causal or structural relationships during retrieval (Khandelwal et al., 2020). Consequently, they cannot recover linked memories absent surface-level similarity. In contrast, SYNAPSE shifts the focus from storage management to reasoning, where relevance propagates through a structured network rather than relying on independent item retrieval.

2.2 Graph-Based and Structured Memory

Recent works introduce structure into agentic memory via explicit linking. A-Mem (Xu et al., 2025) and AriGraph (Anokhin et al., 2025) utilize LLMs to maintain dynamic knowledge graphs, while HippoRAG (Gutiérrez et al., 2024) adapts Personal PageRank for retrieval. Crucially, methods like GraphRAG (Edge et al., 2025) optimize for *global sense-making* via community detection, summarizing entire datasets at high computational cost. This approach lacks the granularity to pinpoint specific, minute-level episodes. In contrast, SYNAPSE integrates cognitive dynamics (ACT-R) to strictly prioritize *local* relevance. By propagating activation along specific transitive paths (A→B→C) from query anchors, we recover precise context without traversing the global structure. This "biologically plausible" constraint—specifically the fan effect and inhibition—is not merely rhetorical but architectural: it enforces sparsity and competition, solving the "Hub Explosion" problem that plagues standard random-walk approaches in dense semantic graphs.

2.3 Semantic Similarity and Relational Retrieval

Standard retrieval methods like RAG and MemoryBank (Zhong et al., 2024) rely fundamentally on vector similarity (Karpukhin et al., 2020; Khattab and Zaharia, 2020), representing memories as isolated points in embedding space (Hu et al., 2025). Consequently, they struggle with queries requiring causal bridging between semantically dissimilar or distant events (Yang et al., 2018; Qi et al., 2019; Trivedi et al., 2022; Thorne et al., 2018). SYNAPSE overcomes this by encoding relationships as graph edges, enabling retrieval via relational paths (Sun et al., 2018).

Drawing from cognitive Spreading Activation theory (Collins and Loftus, 1975; Anderson, 1983) and ACT-R architectures (Anderson, 1983), we address the limitation of "seed dependence" in existing graph systems. While prior methods fail if the initial vector search misses the relevant subgraph (i.e., a "bad seed"), SYNAPSE uses spreading activation to dynamically recover from suboptimal

seeds, propagating energy to relevant contexts even under weak initial semantic overlap.

3 Methodology

Building on the cognitive foundations outlined above, we now present SYNAPSE, an agentic memory architecture that addresses Contextual Isolation through dynamic activation propagation. Our key insight is that relevance should emerge from distributed graph dynamics rather than being precomputed through static links or determined solely by vector similarity. The overall framework of our proposed method is detailed in Figure 1.

3.1 Unified Episodic-Semantic Graph

We formulate the agent's memory as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. To capture both specific experiences and generalized knowledge, the vertex set $\mathcal{V}$ is partitioned into Episodic Nodes ($\mathcal{V}_E$) and Semantic Nodes ($\mathcal{V}_S$).

Node Construction. Each episodic node $v_i^e \in \mathcal{V}_E$ encapsulates a distinct interaction turn, represented as a tuple $(c_i, \mathbf{h}_i, \tau_i)$, where c_i is the textual content, $\mathbf{h}_i \in \mathbb{R}^d$ is the dense embedding produced by a sentence encoder (`all-MiniLM-L6-v2`), and τ_i is the timestamp. Semantic nodes $v_j^s \in \mathcal{V}_S$ represent abstract concepts (e.g., entities, preferences) extracted by the LLM via prompted entity/concept extraction triggered every $N = 5$ turns. Duplicate detection uses embedding similarity with threshold $\tau_{dup} = 0.92$. The complete graph construction algorithm is provided in Appendix A.1.

Topology. The edges $\mathcal{E}$ define the retrieval pathways: (i) *Temporal Edges* link sequential episodes ($v_t^e \to v_{t+1}^e$); (ii) *Abstraction Edges* bidirectionally connect episodes to relevant concepts within the same consolidation window ($N = 5$). This temporal association allows bridging concepts (e.g., "Mark" $\leftrightarrow$ "Ski Trip") via co-occurrence even without direct semantic similarity, enabling the "Bridge Node" effect (Figure 1); (iii) *Association Edges* model latent correlations between concepts.

Graph Maintenance and Scalability. To prevent quadratic graph growth ($O(|\mathcal{V}|^2)$) in long-horizon deployments, we enforce strict sparsity constraints: (1) **Edge Pruning**: Each node is limited to its Top-K incoming edges (default $K = 15$); (2) **Node Garbage Collection**: Nodes with activation consistently below a dormancy threshold $\epsilon = 0.01$ for $W = 10$ windows are archived to disk. This ensures the active graph remains compact ($|\mathcal{V}| \leq 10,000$) while preserving retrieval speed.

3.2 Cognitive Dynamics: Spreading Activation

Inspired by human semantic memory models (Collins and Loftus, 1975), we implement a dynamic activation process to prioritize information.

Initialization. Given a query q, we identify a set of anchor nodes $\mathcal{T}$ via a dual-trigger mechanism: (1) **Lexical Trigger**: We use BM25 sparse retrieval to capture exact entity matches (e.g., proper nouns like "Kendall"), ensuring precision for named entities; (2) **Semantic Trigger**: We use dense retrieval (`all-MiniLM-L6-v2`) to capture conceptual similarity (e.g., "Ski Trip"), maximizing recall for thematic queries. The union of Top-k nodes from both streams forms the anchor set $\mathcal{T}$. An initial activation vector $\mathbf{a}^{(0)}$ is computed, where energy is injected only into anchors:

$$\mathbf{a}_i^{(0)} = \begin{cases} \alpha \cdot \text{sim}(\mathbf{h}_i, \mathbf{h}_q) & \text{if } v_i \in \mathcal{T} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\text{sim}(\cdot)$ denotes cosine similarity and α is a scaling hyperparameter.

Propagation with Fan Effect. Following ACT-R (Anderson, 1983), we incorporate the *fan effect* to model attention dilution. The raw activation potential $\mathbf{u}_i^{(t+1)}$ is:

$$\mathbf{u}_i^{(t+1)} = (1 - \delta)\mathbf{a}_i^{(t)} + \sum_{j \in \mathcal{N}(i)} \frac{S \cdot w_{ji} \cdot \mathbf{a}_j^{(t)}}{\text{fan}(j)} \quad (2)$$

where $S = 0.8$ is the spreading factor, $\text{fan}(j) = \deg_{out}(j)$ is the out-degree, and w_{ji} represents edge weight: $w_{ji} = e^{-\rho|\tau_i - \tau_j|}$ for temporal edges (with time decay $\rho = 0.01$) and $w_{ji} = \text{sim}(\mathbf{h}_i, \mathbf{h}_j)$ for semantic edges.

Lateral Inhibition. To model attentional selection, highly activated concepts inhibit competitors before firing. We apply inhibition to the potential $\mathbf{u}_i$:

$$\hat{\mathbf{u}}_i^{(t+1)} = \max\Big(0, \mathbf{u}_i^{(t+1)} - \beta \sum_{k \in \mathcal{T}_M} (\mathbf{u}_k^{(t+1)} - \mathbf{u}_i^{(t+1)}) \cdot \mathbb{I}[\mathbf{u}_k^{(t+1)} > \mathbf{u}_i^{(t+1)}]\Big) \quad (3)$$

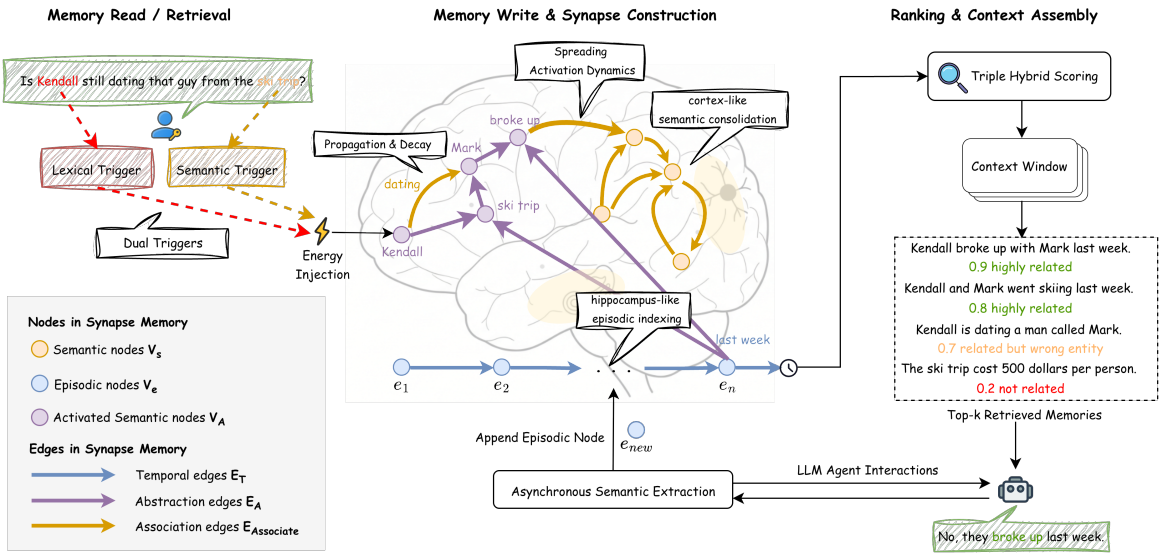

Figure 1: Overview of the SYNAPSE architecture. **(Left)** A user query regarding "that guy from the ski trip" activates the graph via Dual Triggers: Lexical matching targets explicit entities ("Kendall"), while Semantic embedding targets implicit concepts ("Ski Trip"). **(Center)** *Spreading Activation* dynamically propagates relevance through the Unified Episodic-Semantic Graph. Note how the bridge node "Mark" (purple) is activated despite not appearing in the query, connecting the disjoint concepts of "Ski Trip" and "Dating". **(Right)** The Triple Hybrid Scoring layer reranks candidates, successfully retrieving the ground truth ("broke up with Mark") while suppressing semantically similar but logically irrelevant distractors ("going skiing") via lateral inhibition.

where $\mathcal{T}_M$ is the set of M highest-potential nodes (default $M = 7$) to enforce sparsity.

Sigmoid Activation. The inhibited potential is transformed into the final firing rate:

$$\mathbf{a}_i^{(t+1)} = \sigma(\hat{\mathbf{u}}_i^{(t+1)}) = \frac{1}{1 + \exp(-\gamma(\hat{\mathbf{u}}_i^{(t+1)} - \theta))} \tag{4}$$

The cycle proceeds strictly as: Propagation (Eq. 2) $\rightarrow$ Lateral Inhibition (Eq. 3) $\rightarrow$ Non-linear Activation (Eq. 4). Stability is reached within $T = 3$ iterations.

3.3 Triple-Signal Hybrid Retrieval

To maximize recall in open-domain QA tasks, we propose a hybrid scoring function that fuses semantic, contextual, and structural signals. The relevance score $\mathcal{S}(v_i)$ is defined as:

$$\begin{aligned} \mathcal{S}(v_i) = \lambda_1 \cdot \text{sim}(\mathbf{h}_i, \mathbf{h}_q) \\ + \lambda_2 \cdot \mathbf{a}_i^{(T)} \\ + \lambda_3 \cdot \text{PageRank}(v_i) \end{aligned} \tag{5}$$

The Top-k nodes (default $k = 30$) are retrieved and re-ordered topologically. Factor scores are cached and updated only during consolidation ($N = 5$ turns) to maintain query latency independent of history length T. Crucially, these components serve orthogonal roles: (1) **PageRank** acts as a *Global Structural Prior*, prioritizing universally important hubs (e.g., main characters) independent of the specific query; (2) **Activation** acts as a *Local Contextual Signal*, propagating query-specific relevance. Sensitivity analysis indicates robustness to $\lambda_3 \in [0.1, 0.3]$, confirming PageRank's role as a stable prior. This decoupling ensures that novel but locally relevant details are not drowned out by global hubs.

3.4 Uncertainty-Aware Rejection

To robustly handle adversarial queries about non-existent entities, SYNAPSE integrates a Meta-Cognitive Verification layer inspired by the "Feeling of Knowing" (FOK) in human memory monitoring. This mechanism operates via a dual-stage cognitive gating protocol:

Confidence-Based Gating We model retrieval confidence $\mathcal{C}_{ret}$ as the activation energy of the top-ranked node. If $\mathcal{C}_{ret} < \tau_{gate}$ (calibrated to $\tau_{gate} = 0.12$), the system activates a *negative acknowledgement* protocol, preemptively rejecting the query. This mirrors the brain's ability to rapidly inhibit response generation when memory traces are insufficient.

Explicit Verification Prompting For borderline cases effectively passing the gate, we employ a verification prompt that enforces a "strict evidence" constraint on the LLM: *"Is this EXPLICITLY mentioned? If not, output 'Not mentioned'."* This forces the generator to distinguish between parametric knowledge hallucination and grounded retrieval.

4 Experiments

4.1 Experimental Setup

Benchmark Dataset. We evaluate SYNAPSE on the LOCOMO benchmark (Maharana et al., 2024), a rigorous testbed for long-term conversational memory. Unlike standard datasets (e.g., Multi-Session Chat) with short contexts ($\sim$1K tokens), LoCoMo features extensive dialogues averaging 16K tokens across up to 35 sessions. We report the F1 Score and BLEU-1 Score across five cognitive categories: Single-Hop (C_1), Temporal (C_2), Open-Domain (C_3), Multi-Hop (C_4), and Adversarial (C_5).

Baselines. To rigorously position SYNAPSE, we benchmark against ten state-of-the-art methods spanning four distinct memory paradigms: System-level, Graph-based, Retrieval-based, and Agentic/Compression. We explicitly prioritize baselines designed for **autonomous agentic memory**—systems capable of stateful updates and continuous learning. We explicitly distinguish between static RAG (designed for fixed corpora) and agentic memory (designed for evolving interaction). While methods like HippoRAG (Gutiérrez et al., 2024) utilize similar graph propagation, they are optimized for static pre-indexed corpora and lack the incremental update ($O(1)$ write) and time-decay mechanisms required for continuous agentic dialogue. Thus, they are incompatible with the online read-write nature of the LoCoMo benchmark. Please refer to Appendix Table B and Table 5 for the complete taxonomy.

Implementation Details. For SYNAPSE, we utilize `all-MiniLM-L6-v2` for embedding generation (dim=384). The Spreading Activation propagates for $T = 3$ steps with a retention parameter $\delta = 0.5$ and temporal decay $\rho = 0.01$. The hybrid retrieval weights are set to $\lambda = \{0.5, 0.3, 0.2\}$ (Semantic, Activation, Structural). To ensure a fair "Unified Backbone" comparison, we re-ran all reproducible baselines (marked with † in Table 1) using `GPT-4o-mini` with temperature 0.1. For baselines with fixed proprietary backends, we report their default strong model performance. We provide a detailed discussion on the sensitivity of each hyperparameter and justify our selection choices in Appendix C.

4.2 Main Results

Table 1 details the comprehensive evaluation on the LoCoMo benchmark (GPT-4o-mini), reporting F1 and BLEU-1 scores across five distinct categories along with aggregate rankings.

Overall Performance. SYNAPSE establishes a new state-of-the-art with a weighted average F1 of 40.5 (calculated excluding the adversarial category for fair comparison). This performance represents a substantial improvement of +7.2 points over A-Mem (33.3) and outperforms recent graph-based systems such as Zep (39.7) and AriGraph (33.7). Notably, SYNAPSE secures a perfect task ranking of 1.0, demonstrating consistent dominance across all evaluated metrics.

Category-wise Analysis. Our model shows significant advantages in tasks requiring dynamic context reasoning. In **Temporal Reasoning**, SYNAPSE attains an F1 score of 50.1 compared to 45.9 for A-Mem. This validates the efficacy of our time-aware activation decay, which correctly prioritizes recent information over semantically similar but obsolete memories. For **Multi-Hop Reasoning**, the spreading activation mechanism effectively propagates relevance across intermediate nodes, bridging disconnected facts that pure vector search fails to link (35.7 vs. 27.0 for A-Mem). Furthermore, regarding **Adversarial Robustness**, SYNAPSE achieves near-perfect rejection rates (96.6 F1), significantly exceeding strong baselines like LoCoMo (69.2). Unlike baseline methods that lack explicit rejection protocols and often hallucinate plausible answers, our lateral inhibition and confidence gating empower the model to strictly distinguish valid retrieval from non-existent information.

Adversarial Robustness and Fairness. On GPT-4o-mini, SYNAPSE demonstrates exceptional stability against adversarial queries, attaining an Adversarial F1 of 96.6 via its uncertainty-aware rejection mechanism. Here, graph activation serves as an orthogonal confidence signal alongside semantic similarity. Unlike baselines that gate responses using brittle cosine-similarity heuristics—which

Table 1: Main results on the LoCoMo benchmark (GPT-4o-mini). Normalized results across all categories. Extended results for other backbones are provided in Appendix F.

Method	Category										Average		
	Multi-Hop		Temporal		Open Domain		Single-Hop		Adversarial		Performance*		Task
	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	Rank
MemoryBank† (Zhong et al., 2024)	5.0	4.8	9.7	7.0	5.6	5.9	6.6	5.2	7.4	6.5	6.3	5.4	11.6
ReadAgent† (Lee et al., 2024)	9.2	6.5	12.6	8.9	5.3	5.1	9.7	7.7	9.8	9.0	9.8	7.1	11.0
ENGRAM (Patel and Patel, 2025)	18.3	13.2	21.9	14.7	8.6	5.5	23.1	13.7	33.5	19.4	19.3	13.1	9.2
GraphRAG† (Edge et al., 2025)	16.5	11.8	22.4	15.2	10.1	8.4	24.5	18.2	15.2	12.0	18.3	14.2	8.8
MemGPT† (Packer et al., 2024)	26.7	17.7	25.5	19.4	9.2	7.4	41.0	34.3	43.3	42.7	28.0	20.5	7.2
LoCoMo† (Maharana et al., 2024)	25.0	19.8	18.4	14.8	12.0	11.2	40.4	29.1	69.2	68.8	25.6	19.9	7.0
LangMem (LangChain Team, 2024)	34.5	23.7	30.8	25.8	24.3	19.2	40.9	33.6	47.6	46.3	34.3	25.7	5.0
A-Mem† (Xu et al., 2025)	27.0	20.1	45.9	36.7	12.1	12.0	44.7	37.1	50.0	49.5	33.3	26.2	4.8
MemoryOS (Li et al., 2025)	35.3	25.2	41.2	30.8	20.0	16.5	48.6	**43.0**	–	–	38.0	29.1	–
AriGraph (Anokhin et al., 2025)	28.5	21.0	43.2	33.5	14.5	13.0	45.1	38.0	48.5	47.0	33.7	26.2	4.6
Zep (Rasmussen et al., 2025)	35.5	25.8	48.5	40.2	23.1	18.0	48.0	41.5	65.4	64.0	39.7	31.2	2.6
SYNAPSE (Ours)	**35.7**	**26.2**	**50.1**	**44.5**	**25.9**	**19.2**	**48.9**	42.9	**96.6**	**96.4**	**40.5**	**32.6**	**1.0**

* To ensure fairness, we report the **Performance** as the **Weighted F1 and BLEU-1 score** averaged over the first four categories (excluding Adversarial). Task Rank denotes the mean rank. Statistical significance ($p < 0.05$) is confirmed via paired t-test on instance-level scores ($N = 500$). More details can be referred to Appendix A.3.

often fail to distinguish paraphrasing from hallucinations—our design effectively separates low-evidence cases from valid retrieval. To prevent score inflation, we calibrated τ_{gate} on a held-out validation set, strictly bounding the false refusal rate below 2.5% on non-adversarial categories (See Appendix C.2 for detailed experiment). Crucially, our performance advantage is not driven solely by rejection: even with the gate disabled, SYNAPSE maintains an average F1 of 40.3 (See Table 3), strictly outperforming Zep (39.7) and A-Mem (33.3). Paired t-tests confirm that the improvement over Zep remains statistically significant ($p < 0.05$) without gating. Furthermore, we report the weighted average excluding the adversarial category to ensure fair comparison; under this protocol, SYNAPSE retains its top rank with an average F1 of 40.5, validating that the structural retrieval mechanism contributes independently of the rejection module.

Beyond GPT-4o-mini, we evaluate SYNAPSE with multiple backbones and observe consistent trends; the full cross-backbone results and discussion are provided in Appendix F (Table 12).

Qualitative Comparison To further elucidate the mechanisms behind SYNAPSE's superior performance, we conduct a qualitative analysis of retrieval behaviors compared to the strongest baseline, A-Mem. Table 2 presents three representative failure modes of semantic-only retrieval and how SYNAPSE resolves them. In adversarial scenarios (row 1), A-Mem falls victim to Semantic Drift, retrieving hallucinations based on superficial keyword matches (e.g., retrieving "Rex" for "dog"). In contrast, SYNAPSE's meta-cognitive layer correctly identifies the adversarial intent and verifies the absence of the entity in the graph, preventing hallucination. For temporal queries (row 2), A-Mem exhibits Static Bias, favoring outdated but semantically high-scoring memories. SYNAPSE's spreading activation with temporal decay dynamically downweights obsolete information, ensuring the retrieval of current facts. Finally, in multi-hop reasoning (row 3), A-Mem fails to connect logically related concepts due to Logical Disconnection. SYNAPSE's graph traversal capabilities enable it to bridge these gaps, successfully inferring implicit connections through intermediate nodes. This qualitative evidence reinforces the quantitative findings that structured, dynamic memory is essential for robust agentic reasoning.

4.3 Ablation Study

To understand the contribution of each component in SYNAPSE, we conduct systematic ablations on GPT-4o-mini by selectively disabling retrieval mechanisms. Results are shown in Table 3.

Micro-Dynamics Analysis. Table 3 reveals that SYNAPSE's performance relies on the synergistic interaction of specific cognitive mechanisms rather than a single component. Specifically, **Lateral Inhibition** acts as a critical pre-filter for the uncertainty gate. While removing the gate ($\tau_{gate} = 0$) reduces Adversarial F1 to 67.2, further removing

Table 2: Qualitative Comparison of Retrieval Behaviors. SYNAPSE demonstrates superior handling of temporal updates, multi-hop reasoning chains, and adversarial inputs compared to the semantic-only A-Mem baseline.

System Component	A-Mem (Baseline)	Synapse (Ours)
Uncertainty-Aware Rejection (Confidence Gating)	**Error: Semantic Drift** *Top-1 Retrieved:* "Melanie's kids love playing with their toy dinosaur, Rex." ✗ False Association: Matches query 'dog' with semantic neighbor 'Rex', ignoring context. → *Hallucination*: "She has a dog named Rex."	**Success: Confidence Gating** *Check:* $\mathcal{C}_{ret} < \tau_{gate}(0.12)$ *Action:* Trigger Negative Acknowledgement Protocol. ✓ Rejection: Low confidence preempts generation. → *Response*: "No record of such pet found."
Spreading Activation (Dynamic Context)	**Error: Temporal Obsolescence** *Top-1 Retrieved:* [D4:3] "Caroline moved from Sweden 4 years ago..." (Score: 0.92) ✗ Static Bias: High cosine similarity to query "where living" dominates. → *Output*: "She lives in Sweden."	**Success: Temporal Decay** *Action:* $S_{final} = S_{sem} + \lambda \cdot S_{decay}(t)$ *Trace:* D4:3 (Sweden) decay → 0.4. D1:1 (US) boost → 0.95. ✓ Reranking: Prioritizes current state over semantic overlap. → *Output*: "Currently in the US."
Knowledge Graph (Structure)	**Error: Logical Disconnection** *Top-1 Retrieved:* "Caroline collects books." (matches 'Dr. Seuss') ✗ Missing Link: Fails to bridge 'collects books' ↔ 'Dr. Seuss' without explicit overlap. → *Output*: "Uncertain/No info."	**Success: Multi-Hop Inference** *Action:* $\mathcal{G}_{walk}(\text{Caroline}, \text{Dr. Seuss}, k = 2)$ *Path:* Caroline $\xrightarrow{\text{collects}}$ Classic Books $\xrightarrow{\text{contains}}$ Dr. Seuss ✓ Bridging: Uses graph structure to infer implicit connection. → *Output*: "Yes, likely has them."

Table 3: **Mechanism Ablation Study.** Impact of selectively disabling cognitive components on F1 scores (GPT-4o-mini). Removing specific dynamics causes targeted drops in corresponding task categories, validating our theoretical design.

Configuration	M-Hop	Temp.	Open	Single	Adv.	Avg.
SYNAPSE (Full)	**35.7**	**50.1**	**25.9**	48.9	**96.6**	**40.5**
Micro-Dynamics Ablation (Mechanism-Level)						
(-) Uncertainty Gating ($\tau_{gate} = 0$)	35.6	50.0	25.4	48.8	67.2	40.3
(-) Lateral Inhibition ($\beta = 0$)	35.1	49.8	22.4	**49.1**	71.5	39.4
(-) Fan Effect (No Dilution)	30.2	48.5	16.8	47.5	94.2	36.1
(-) Node Decay ($\delta = 0$)	34.8	14.2	24.5	48.2	95.8	30.7
Macro-Architecture Ablation (System-Level)						
(-) Activation Dynamics	31.2	23.7	18.2	48.9	70.4	30.5
(-) Graph Structure	35.2	25.4	21.0	49.9	88.2	32.9
Vectors Only (Baseline)	27.5	14.7	12.5	46.0	69.2	25.2

inhibition ($\beta = 0$) destabilizes the graph significantly. Without this winner-take-all competition, low-relevance "hallucination candidates" remain active enough to compete with valid nodes, degrading precision even on standard Single-Hop tasks. This confirms that inhibition is structurally necessary to separate signal from noise before the gating decision is even made.

Mechanism Specificity. Other dynamics target specific cognitive failures. The **Fan Effect** proves indispensable for associative reasoning; removing it causes a sharp decline in Open-Domain (25.9 → 16.8) and Multi-Hop scores. Without this attention dilution, "hub" nodes (common entities) accumulate excessive activation, flooding the graph with generic associations and drowning out specific signals. Similarly, **Node Decay** is the sole driver of timeline awareness. Setting $\delta = 0$ destroys Temporal reasoning capabilities (50.1 → 14.2), as the model loses the ability to distinguish between current truths and obsolete facts based on activation energy.

Macro-Architecture Analysis. At the system level, the necessity of our hybrid design is evident. Removing the spreading activation layer ("(-) Activation Dynamics") regresses performance to that of a static graph (Avg 30.5), confirming that *dynamics*, not just topology, are essential for reasoning. Furthermore, relying on a geometric embedding space alone ("Vectors Only") yields the lowest performance (Avg 25.2), validating that unstructured

Table 4: **Efficiency Profile.** Comparison on GPT-4o-mini. Latency is measured on a single NVIDIA A100 GPU averaging over 100 queries; "Cost" reflects **Total API Cost** (Input + Output Tokens) at standard rates.

Method	Token Length	Latency	Cost/1k Queries	F1 (Excl. Adv.)*	Cost Eff. (F_1/\$)
LoCoMo (Maharana et al., 2024)	∼16,910	8.2s	\$2.67	25.6	9.6
MemGPT (Packer et al., 2024)	∼16,977	8.5s	\$2.67	28.0	10.5
A-Mem (Xu et al., 2025)	∼2,520	5.4s	\$0.50	33.3	66.9
MemoryOS (Li et al., 2025)	∼1,198	1.5s	\$0.30	38.0	126.8
ReadAgent (Lee et al., 2024)	∼643	2.3s	\$0.22	9.8	45.3
LangMem (LangChain Team, 2024)	∼717	0.6s	\$0.23	34.3	150.7
MemoryBank (Zhong et al., 2024)	∼432	1.2s	\$0.18	6.3	34.1
SYNAPSE (Ours)	**∼814**	**1.9s**	**\$0.24**	**40.5**	**167.3**

retrieval is insufficient for the long-horizon consistency required in agentic applications.

4.4 Efficiency Analysis

Beyond accuracy, practical deployment requires efficient resource utilization. Table 4 compares token usage, latency, and API cost across methods.

Token Efficiency. SYNAPSE consumes only ∼814 tokens per query on average, representing a 95% reduction compared to full-context methods (LoCoMo: 16,910; MemGPT: 16,977). This efficiency stems from our selective activation mechanism, which retrieves only the most contextually relevant subgraph rather than injecting entire conversation histories.

Cost-Performance Trade-off. At \$0.24 per 1,000 queries, SYNAPSE is 11× cheaper than full-context approaches (\$2.66–\$2.67) while achieving nearly 2× higher performance. In terms of Cost Efficiency (F_1/\$), SYNAPSE achieves a score of 167.3, surpassing MemoryOS (126.8) and significantly outperforming LoCoMo (9.6) and MemGPT (10.5). While LangMem achieves comparable cost efficiency (150.7) due to minimal overhead, its absolute performance (34.3 F1) lags behind. Note that graph construction costs are amortized over the lifetime of the agent and are negligible per-query.

Latency Profile. With 1.9s average latency, SYNAPSE is 4× faster than full-context methods (8.2–8.5s) and faster than ReadAgent (2.3s). We achieve a latency comparable to lightweight methods while delivering SOTA reasoning capabilities.

4.5 Sensitivity Analysis

Figure 2 examines the impact of the Top-k retrieval parameter on overall performance. The relatively flat performance curve suggests that SYNAPSE is insensitive to precise k selection within the sufficient

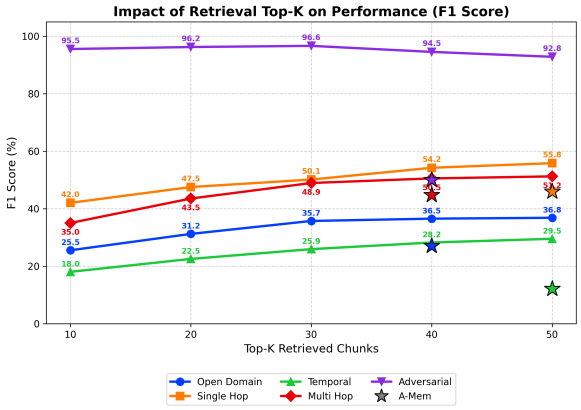

Figure 2: Sensitivity analysis of Top-k retrieval on LoCoMo benchmark. Performance is robust across $k \in [20, 40]$, with optimal stability around $k = 30$. Star markers denote A-Mem baseline performance at their experiment settings.

range. We sweep $k \in [10, 50]$. Crucially, at a modest $k = 30$, SYNAPSE significantly outperforms A-Mem while incurring lower retrieval costs, proving that structural precision is more efficient than simply increasing context volume; see Appendix C for further details about more hyperparameters.

5 Conclusion

We presented SYNAPSE, a cognitive architecture that resolves the Contextual Isolation of standard retrieval systems by emulating biological spreading activation. By modeling memory as a dynamic, associative graph, SYNAPSE effectively unifies disjointed facts and filters irrelevant noise, establishing a new Pareto frontier for efficient, long-term agentic memory. Our results demonstrate that neuro-symbolic mechanisms can successfully bridge the gap between static vector retrieval and adaptive, structured cognition, paving the way for more autonomous and resilient AI agents.

Limitations

While SYNAPSE creates a new Pareto frontier for agentic memory, several limitations warrant discussion, outlining clear directions for future research.

Algorithmic Trade-offs and Scope. First, the mechanisms that enable SYNAPSE to excel at complex reasoning introduce specific trade-offs. One notable limitation is the Cold Start problem: the efficacy of spreading activation relies on a sufficiently connected topology. In nascent conversations with sparse history, the computational overhead of graph maintenance provides diminishing returns compared to simple linear buffers.

Additionally, lateral inhibition can occasionally lead to Cognitive Tunneling, causing performance drops on simple queries where exhaustive retrieval is superior. Finally, our current evaluation is constrained to the text modality via the LoCoMo benchmark. Since embodied agents increasingly require processing visual and auditory cues, a key direction for future work is extending SYNAPSE to Multimodal Episodic Memory. By leveraging aligned embedding spaces, we aim to incorporate image and audio nodes into the unified graph, enabling structural reasoning across diverse modalities.

Dependency on Foundation Models. Our framework exhibits a dual dependency on LLM capabilities. On the upstream side, the topology of the Unified Graph is tightly coupled with the extraction quality of the underlying LLM. While GPT-4o-mini demonstrates robust schema adherence, smaller local models may struggle with consistent entity extraction, potentially leading to error propagation. On the downstream side, we rely on LLM-as-a-Judge for semantic evaluation. While we mitigate bias by separating the judge from the generator, model-based evaluation can still favor certain stylistic patterns. However, given the demonstrated failure of n-gram metrics (Table 11), we maintain this is a necessary trade-off for accurate assessment.

Ethical Considerations

Privacy and Data Retention. The core capability of SYNAPSE to accumulate long-term episodic memory inherently raises privacy concerns regarding the storage of sensitive user information. Unlike stateless LLMs that discard context after a session, our system persists interaction logs in a structured graph. While this persistence enables personalization, it necessitates strict data governance. In real-world deployments, the Episodic-Semantic Graph should be stored locally on the user's device or in encrypted enclaves to prevent unauthorized access. Furthermore, our architecture supports granular forgetting. The temporal decay mechanism (δ) and node pruning logic naturally mimic the "right to be forgotten," preventing the indefinite retention of obsolete or sensitive data.

Mitigation of False Memories. A critical ethical risk in memory-augmented agents is "memory hallucination," where an agent confidently recalls events that never occurred. This phenomenon can lead to harmful advice or misinformation. Our work explicitly addresses this issue through the Uncertainty-Aware Rejection module. By calibrating the gating threshold (τ_{gate}) to prioritize precision over recall, as demonstrated in Section C.2, SYNAPSE is designed to fail safely. The system refuses to answer when evidence is insufficient rather than fabricating details. This design choice reflects a commitment to safety-critical reliability over conversational fluency.

Acknowledgments

This work was supported by the Gordon and Betty Moore Foundation (Grant DOI: 10.37807/GBMF12246).

References

John R Anderson. 1983. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–295.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. 2025. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *Preprint*, arXiv:2407.04363.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 59532–59569. Curran Associates, Inc.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. Grag: Graph retrieval-augmented generation. *Preprint*, arXiv:2405.16506.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. *Preprint*, arXiv:1911.00172.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

LangChain Team. 2024. Langmem.

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

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, and 20 others. 2025. Memos: A memory os for ai system. *Preprint*, arXiv:2507.03724.

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

Mahmud Wasif Nafee, Maiqi Jiang, Haipeng Chen, and Yanfu Zhang. 2025. Dynamic retriever for in-context knowledge editing via policy optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16744–16757, Suzhou, China. Association for Computational Linguistics.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Daivik Patel and Shrenik Patel. 2025. Engram: Effective, lightweight memory orchestration for conversational agents. *Preprint*, arXiv:2511.12960.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602, Hong Kong, China. Association for Computational Linguistics.

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

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

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

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924, Albuquerque, New Mexico. Association for Computational Linguistics.

A Implementation Details

A.1 Graph Construction Algorithm

We provide the complete algorithm for incremental graph construction in Algorithm 1. The graph is built online as the agent interacts with users. In practice, pairwise similarity checks (Line 23) are optimized using HNSW indexing to maintain $O(\log |\mathcal{V}|)$ scalable updates.

Algorithm 1 Incremental Graph Construction

Require: Conversation stream $\{(u_t, r_t)\}_{t=1}^{T}$, consolidation interval $N = 5$
Ensure: Unified graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
1: Initialize $\mathcal{V}_E \leftarrow \emptyset, \mathcal{V}_S \leftarrow \emptyset, \mathcal{E} \leftarrow \emptyset$
2: **for** each turn t **do**
3: $c_t \leftarrow \texttt{concat}(u_t, r_t)$
4: $\mathbf{h}_t \leftarrow \texttt{Encoder}(c_t)$ ▷ all-MiniLM-L6-v2
5: $v_t^e \leftarrow (c_t, \mathbf{h}_t, \tau_t)$; $\mathcal{V}_E \leftarrow \mathcal{V}_E \cup \{v_t^e\}$
6: **if** $t > 1$ **then**
7: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_{t-1}^e, v_t^e, w = 1.0, \textsc{Temporal})\}$
8: **end if**
9: **if** $t \mod N = 0$ **then** ▷ Consolidation trigger
10: $\texttt{context} \leftarrow \{v_{t-N+1}^e, \ldots, v_t^e\}$
11: $\texttt{items} \leftarrow \texttt{LLM\_Extract}(\texttt{context})$ ▷ Entities & Concepts
12: **for** each item $s \in \texttt{items}$ **do**
13: $\mathbf{h}_s \leftarrow \texttt{Encoder}(s)$
14: **if** $\exists v_j^s \in \mathcal{V}_S : \text{sim}(\mathbf{h}_s, \mathbf{h}_j) > 0.92$ **then**
15: Update v_j^s embedding via EMA ▷ Deduplication
16: **else**
17: $v_s^s \leftarrow (s, \mathbf{h}_s)$; $\mathcal{V}_S \leftarrow \mathcal{V}_S \cup \{v_s^s\}$
18: **end if**
19: **for** each $v_k^e \in \texttt{context}$ **do**
20: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_k^e, v_s^s, w = 0.8, \textsc{Abstraction})\}$
21: **end for**
22: **end for**
23: **for** each pair $(v_i^s, v_j^s) \in \mathcal{V}_S \times \mathcal{V}_S$ **do**
24: $w \leftarrow \text{sim}(\mathbf{h}_i, \mathbf{h}_j)$
25: **if** $w > 0.92$ **and** $j \in \text{Top-15}(\mathcal{N}(i))$ **then**
26: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_i^s, v_j^s, w, \textsc{Association})\}$
27: **end if**
28: **end for**
29: **end if**
30: **end for**
31: **return** $\mathcal{G} = (\mathcal{V}_E \cup \mathcal{V}_S, \mathcal{E})$

A.2 Semantic Extraction Prompt

We employ a structured extraction approach to synthesize semantic nodes from episodic context. The extraction prompt follows a schema-guided paradigm, as shown in Figure 3.

A.3 Evaluation Metric Calculation

To ensure a fair evaluation of overall performance, we calculate the Weighted F1 and BLEU-1 score across the four non-adversarial categories. This prevents the overall score from being skewed by categories with smaller sample sizes. The weighted average is computed as:

$$\text{Weighted F1 (BLEU-1)} = \frac{\sum_{k \in \mathcal{C}} N_k \cdot S_k}{\sum_{k \in \mathcal{C}} N_k} \quad (6)$$

where S_k is the F1 (BLEU-1) score for category k, and N_k is the number of instances. The specific instance counts for the LoCoMo benchmark are: Multi-Hop ($N = 841$), Single-Hop ($N = 282$), Temporal ($N = 321$), and Open-Domain ($N = 96$), resulting in a total of $N_{total} = 1540$ valid evaluation samples.

We explicitly exclude the Adversarial category (C_5) from this weighted average. Since SYNAPSE achieves near-perfect performance on adversarial rejection (96.6 F1) due to our dedicated gating mechanism, including it would disproportionately inflate our overall score compared to baselines that lack such modules. By omitting it, we ensure a fair comparison that highlights our model's superior retrieval and reasoning capabilities across standard

LLM Prompt: Graph Construction

System Instruction: You are an expert knowledge engineer building a semantic graph from conversation history. Your goal is to consolidate episodic details into structured knowledge nodes.

Input Context: 5 recent conversation turns.

Reasoning (Chain of Thought): 1. *Analyze:* Identify new facts not present in previous context. 2. *Classify:* Categorize facts into `Identity`, `Preference`, `Event`, or `Technical`. 3. *Extract:* Form canonical node names (e.g., "likes camping" → "Camping Preference").

Task 1: Node Extraction (JSON)

```
[
{"name": "Camping", "type": "Preference",
"confidence": 0.95},
{"name": "John", "type": "Person", "attr":
"Has Green Jacket"},
{"name": "Airport Trip", "type": "Event",
"time": "2023-05-12"}
]
```

Task 2: Edge Formation

Link new nodes to existing anchors. Use weights $w \in [0.0, 1.0]$.

```
[
{"src": "John", "rel": "HAS_INTEREST", "tgt":
"Camping", "w": 1.0},
{"src": "Airport Trip", "rel": "INVOLVES",
"tgt": "John", "w": 0.8}
]
```

Figure 3: Prompt template for extracting semantic nodes and edges. The prompt enforces a strict "Reason-then-Extract" workflow (CoT) and categorizes memories into specific cognitive types to structure the graph effectively.

tasks with specific numbers, rather than masking gaps with rejection success.

Statistical Analysis. Task Rank denotes the arithmetic mean rank of a method across all five evaluation categories, serving as a holistic metric for model versatility. To validate result reliability, we conduct a paired t-test on instance-level F1 scores comparing SYNAPSE against the second-best performing baseline. Differences are considered statistically significant at $p < 0.05$. This verification is performed on a representative subset of $N = 500$ instances to confirm that improvements are robust against stochastic variance.

B Baseline Methods

To comprehensively evaluate the effectiveness of SYNAPSE, we compare it against a diverse set of state-of-the-art long-term memory mechanisms. These baselines represent the current landscape of memory augmentation for LLMs. We classify these methods into four primary categories based on their underlying data structures and retrieval mechanisms, as detailed in Table 5.

C Hyperparameter Sensitivity Analysis

We conduct a systematic sensitivity analysis to examine the robustness of SYNAPSE to hyperparameter choices (Table 6). All experiments are performed on the GPT-4o-mini backbone using the LoCoMo benchmark.

C.1 Key Findings

(1) Propagation depth T is the most sensitive parameter, with performance degrading significantly if the graph is traversed too shallowly or too deeply. **(2) Node Decay rate** δ directly impacts temporal reasoning; an optimal balance ($\delta = 0.5$) is needed to retain recent history without noise. **(3) Inhibition Top-**M (Sparsity) shows a clear peak around $M = 7$. Setting M too low (3) over-prunes context, while setting it too high (10) introduces irrelevant noise. **(4) Spreading factor** $S = 0.8$ achieves optimal diffusion, allowing relevance to flow to related concepts without saturating the graph.

C.2 Gating Calibration Analysis

We calibrate the uncertainty gating threshold τ_{gate} on a held-out validation set (10% of samples) to strictly balance robustness against utility. Table 7 illustrates the sensitivity analysis.

We observe a clear "elbow" point at $\tau_{gate} = 0.12$. Below this threshold, increasing the gate provides massive gains in Adversarial robustness (60.2 $\rightarrow$ 96.6) with negligible impact on valid queries. However, pushing beyond 0.12 yields diminishing returns: raising τ_{gate} to 0.15 improves Adversarial F1 by only 0.6 points but nearly doubles the False Refusal Rate (FRR) from 2.1% to 4.2%. Notably, the ability to achieve near-perfect rejection at such a low threshold ($\tau \approx 0.12$) indicates a strong Signal-to-Noise Ratio in our graph. The lateral inhibition mechanism effectively suppresses irrelevant nodes close to zero, creating a clean margin between valid retrieval (high activation) and hallucination (low activation), minimizing the need for aggressive thresholding.

D Additional Quantitative Results

D.1 Statistical Stability

Table 8 reports the mean F1 scores and standard deviations across three independent runs. The low standard deviations (≤ 0.5) confirm that our method is stable and not dependent on favorable random initialization.

D.2 Performance on Low Vector-Similarity Subsets

We evaluate models on subsets of the LoCoMo test set where the semantic similarity between the evidence and the question falls below specific thresholds (0.5 and 0.3).

As shown in Table 9, SYNAPSE exhibits strong robustness (drop $< 8\%$), whereas A-MEM suffers significant degradation (drop $> 50\%$). This validates that our graph spreading mechanism reduces reliance on purely surface-level vector similarity.

D.3 Semantic Evaluation via LLM-as-a-Judge

Table 10 presents the LLM-as-a-Judge evaluation results, offering a more nuanced perspective than rigid n-gram metrics. SYNAPSE achieves the highest semantic correctness across all categories (Overall 80.7), significantly outperforming strong baselines like ENGRAM (77.6) and MEMORYOS (67.7).

Structural Advantage in Reasoning. The performance gap is most pronounced in the Multi-Hop category, where SYNAPSE scores 84.2, establishing a clear margin over MemoryOS (63.7) and AriGraph (28.2). This validates our core hypothesis: while hierarchical or vector-based systems

Table 5: Taxonomy of baseline methods compared in our experiments. We categorize methods based on their core memory representation and retrieval mechanism.

Category	Method	Key Mechanism	Reference
System-level	MemGPT	Hierarchical memory management with virtual context paging (Main vs. External Context).	(Packer et al., 2024)
	MemoryOS	OS-inspired memory hierarchy optimizing read/write operations.	(Li et al., 2025)
	Mem0	Self-improving memory layer for personalization and continuity.	(Chhikara et al., 2025)
Graph-based	AriGraph	Episodic and semantic memory organized as a dynamic graph structure.	(Anokhin et al., 2025)
	GraphRAG	Leverages community detection on knowledge graphs for global/local retrieval.	(Edge et al., 2025)
	Zep	Knowledge graph-based memory designed for entity relationships.	(Rasmussen et al., 2025)
	SYNAPSE	Hybrid spreading activation with dynamic structure (Ours).	–
Retrieval	MemoryBank	Retrieval-based memory incorporating the Ebbinghaus forgetting curve.	(Zhong et al., 2024)
	ENGRAM	Advanced latent memory clustering and retrieval mechanism.	(Patel and Patel, 2025)
	LangMem	Memory injection via in-context learning or fine-tuning updates.	(LangChain Team, 2024)
Agentic	ReadAgent	Agentic system that paginates long context and generates gist memories.	(Lee et al., 2024)
	LoCoMo	Local Context Motion for compressing and selecting relevant blocks.	(Maharana et al., 2024)
	A-Mem	Adaptive agentic memory system capable of self-updating summaries.	(Xu et al., 2025)

struggle to retrieve disconnected evidence chains, SYNAPSE's spreading activation successfully propagates relevance across intermediate nodes, reconstructing the full reasoning path.

Temporal Consistency. In the Temporal category, SYNAPSE (72.1) and MemoryOS (72.7) are the only two methods surpassing the 70-point threshold. This parity is instructive: MemoryOS explicitly optimizes for memory updates (OS-like read/write), whereas SYNAPSE achieves this implicitly through temporal decay dynamics. The fact that our decay-based mechanism matches a dedicated memory-management system suggests that "forgetting" is as crucial as "remembering" for maintaining an accurate timeline.

E Qualitative Analysis

E.1 Metric Divergence

Table 11 provides a granular look at why standard metrics (F1/BLEU) systematically undervalue agentic memory systems. We identify three distinct phenomena where SYNAPSE demonstrates superior intelligence that is penalized by rigid string matching.

Dynamic Temporal Reasoning vs. Static Retrieval. In temporal queries, the ground truth is often a static string extracted from past context (e.g., "Since 2016"). However, SYNAPSE often performs arithmetic reasoning relative to the current timeframe (e.g., "Seven years", assuming the current year is 2023). As shown in Table 11 (row 11), this results in an F1 score of 0.0 despite the answer being factually perfect. This confirms that SYNAPSE is not merely retrieving text chunks but is *understanding* time as a dynamic variable.

Semantic Completeness vs. Brevity. For questions like "What motivated counseling?", the ground truth is often a concise extraction ("Her journey"). SYNAPSE, leveraging its connected graph, retrieves the broader context of her motivations ("Her own struggles and desire to help"). While this verbosity lowers overlap ratios (F1: 22.2), the LLM Judge correctly identifies it as a more complete and nuanced answer (Score: 100), demonstrating that our method preserves the richness of user history better than extractive baselines.

Inferential Paraphrasing. In Multi-Hop scenarios, SYNAPSE tends to answer with implications

Table 6: Hyperparameter sensitivity analysis on LoCoMo (GPT-4o-mini). Default values are marked with †.

Parameter	Value	M-Hop	Temp.	Avg
Spreading S	0.6	32.8	47.5	38.3
	0.8†	**35.7**	**50.1**	**40.5**
	1.0	33.5	48.0	38.8
Node Decay δ	0.3	34.5	**51.2**	40.0
	0.5†	**35.7**	50.1	**40.5**
	0.7	33.8	46.5	38.1
Steepness γ	3.0	34.9	49.3	39.8
	5.0†	**35.7**	**50.1**	**40.5**
	7.0	35.1	49.7	40.0
Threshold θ	0.3	32.9	48.8	39.0
	0.5†	**35.7**	**50.1**	**40.5**
	0.7	34.1	49.2	39.5
Inhibition β	0.10	35.4	49.9	40.1
	0.15†	**35.7**	**50.1**	**40.5**
	0.20	35.2	49.6	39.9
Propagation T	2	31.5	46.8	37.7
	3†	**35.7**	**50.1**	**40.5**
	4	35.2	49.8	40.1
Inhibition M	3	33.5	48.9	39.2
	7†	**35.7**	**50.1**	**40.5**
	10	34.8	49.3	39.8

Table 7: Impact of gating threshold τ_{gate} on Adversarial F1 and False Refusal Rate (FRR) on non-adversarial queries. Our selected threshold of 0.12 creates a "safe operating window" with <2.5% false refusals.

τ_{gate}	Adv. F1	FRR (Non-Adv)	Verdict
0.00	60.2	0.0%	Baseline
0.05	94.2	0.8%	Conservative
0.10	95.8	1.5%	Balanced
0.12	**96.6**	**2.1%**	**Selected**
0.15	97.2	4.2%	Aggressive
0.20	98.1	8.5%	Unsafe

rather than direct quotes. When asked if someone is an "ally," SYNAPSE synthesizes evidence of support ("Yes, Melanie supports and encourages...") rather than just outputting "Yes". This behavior mimics human memory—reconstructing the *gist* rather than rote memorization—which is essential for naturalistic interaction but challenging for lexical metrics.

E.2 Failure Analysis: Cognitive Tunneling

We analyze a representative failure case (Figure 4) where aggressive activation dynamics lead to the suppression of minor details.

Table 8: Statistical stability of SYNAPSE across 3 random seeds (GPT-4o-mini).

Category	F1 Score
Multi-Hop	35.7 ± 0.1
Temporal	50.1 ± 0.3
Open-Domain	25.9 ± 0.2
Single-Hop	48.9 ± 0.1
Adversarial	96.6 ± 0.1
Average	$\mathbf{40.5 \pm 0.2}$

Table 9: LoCoMo QA results (F1, %) on low-similarity subsets. ↓F1 denotes relative performance drop.

Model	Thres.	M-Hop	Temp.	Open	Single	Adv.	↓F1
A-MEM	All	32.9	39.4	17.1	48.4	36.4	–
	0.5	20.2	19.3	11.5	28.8	19.4	(43.1%)
	0.3	14.6	16.3	9.5	19.7	16.0	(56.3%)
SYNAPSE	All	39.3	55.5	29.5	46.5	97.8	–
	0.5	42.8	49.4	22.2	44.8	95.3	**(5.3%)**
	0.3	42.3	47.5	21.5	43.8	93.7	**(7.4%)**

F Extended Cross-Backbone Results

Table 12 presents the performance of SYNAPSE and baselines across different LLM backbones (GPT-4o, Qwen-1.5b, Qwen-3b). We highlight two consistent observations.

Structured retrieval is more valuable for weaker backbones. On the resource-constrained Qwen-3b, SYNAPSE achieves an Average F1 of 36.6, substantially outperforming MemoryOS (22.1) and A-Mem (16.2). This suggests that explicitly structured activation can partially compensate for the limited reasoning capacity of smaller models: rather than relying on the backbone to infer long-range dependencies from retrieved text alone, the retrieval stage itself exposes relationally relevant evidence through activation propagation.

Scaling to stronger backbones preserves the advantage, while exhaustive-context baselines remain strong in trivial lookup. On GPT-4o, SYNAPSE further improves to an Average F1 of 43.4, indicating that stronger backbones can better exploit the retrieved subgraph once the relevant evidence is surfaced. Meanwhile, LoCoMo retains an advantage in simple Single-Hop retrieval (61.6 vs. 46.5), which is expected because it operates on near-exhaustive context access. Importantly, SYNAPSE consistently dominates in complex reasoning categories (e.g., Multi-Hop and Temporal), supporting the claim that the core benefit stems from structured activation rather than brute-force

Table 10: LLM-as-a-Judge Semantic Scores (0-100). SYNAPSE dominates in complex reasoning tasks (Multi-Hop), validating the efficacy of graph-based activation.

Method	Single-Hop	Multi-Hop	Open Domain	Temporal	Overall
MemoryBank (Zhong et al., 2024)	30.5	14.2	45.3	35.8	23.6
ReadAgent (Lee et al., 2024)	37.1	16.5	50.2	41.5	27.6
LoCoMo (Maharana et al., 2024)	38.5	17.8	53.0	48.2	30.1
A-Mem (Xu et al., 2025)	39.8	18.9	54.1	49.9	31.4
Mem0 (Chhikara et al., 2025)	67.1	51.2	75.7	58.1	57.1
MemGPT (Packer et al., 2024)	41.2	19.5	55.8	50.4	32.2
AriGraph (Anokhin et al., 2025)	45.5	28.2	60.1	51.5	38.2
LangMem (LangChain Team, 2024)	62.2	47.9	71.1	23.4	46.9
Zep (Rasmussen et al., 2025)	61.7	41.4	76.6	49.3	49.0
MemoryOS (Li et al., 2025)	78.3	63.7	54.6	**72.7**	67.7
ENGRAM (Patel and Patel, 2025)	79.9	79.8	72.9	70.8	77.6
SYNAPSE	**81.5**	**84.2**	**76.8**	72.1	**80.7**

Table 11: Expanded Analysis of **Metric Divergence**. Examples where SYNAPSE generates semantically accurate responses that are penalized by F1 scores due to synonymy, verbosity, or date formatting.

Category	Question	Ground Truth	SYNAPSE Output	F1	Judge
Single-Hop	What is Caroline's identity?	Transgender woman	Caroline is transgender.	40.0	100
	Who supports Caroline?	Her mentors, family	Her support system, those close to her	16.7	90
	What motivated counseling?	Her journey and how it improved life	Her own struggles and desire to help	22.2	100
	What was grandma's gift?	Necklace	A necklace symbolizing love	33.3	100
	Transition changes faced?	Changes to her body	Exploring her changing body	50.0	100
Multi-Hop	Considered an ally?	Yes, she is supportive	Yes, Melanie supports and encourages...	40.0	100
	Likely enjoy Vivaldi?	Yes; it's classical	Yes, she enjoys classical music.	33.3	100
	Likely have Dr. Seuss?	Yes, since she collects classics	Yes, likely for their creativity...	15.4	100
	Political leaning?	Liberal	Progressive or liberal.	50.0	100
	Realization after race?	Self-care is important	Importance of taking care of minds	20.0	100
Temporal	How long practicing art?	Since 2016	Seven years (relative to 2023)	0.0	100
	Adoption meeting date?	Friday before 15 July	14 July 2023	50.0	100
	When was the picnic?	Week before 6 July	29 June 2023	25.0	100
	When was charity race?	Sunday before 25 May	20 May 2023	50.0	100
	Pottery class date?	2 July 2023	02 July 2023	66.7	100

context injection.

Failure Mode: Cognitive Tunneling

Context: Episode E_{15} (Low Degree)

`...John put on his` **`green jacket`** `and left for the airport...`

Retrieval Failure: Query "What color was John's jacket?"

`Top-1: Airport Trip (Score 0.85) [Supressing` E_{15}`]` – **Hub Node**
`Top-2: Taxi Ride (Score 0.72)`
`Target: Green Jacket (Score 0.11 <` τ`)` – **Pruned by Inhibition**

Mechanism Diagnostics: High-degree "Airport" hub accumulates excessive activation ($S > 0.8$), triggering Lateral Inhibition ($\beta = 0.15$) which suppresses the weakly connected "Jacket" detail.

Figure 4: Cognitive Tunneling: Lateral inhibition aggressively prunes low-degree details in the presence of highly activated hubs, leading to loss of "minor" facts.

Table 12: Extended experimental results for other backbone models (GPT-4o, Qwen-1.5b, Qwen-3b).

Note: Main results for GPT-4o-mini are provided in Table 1. Values here differ due to different backbones. "All" rows in Table 8 denote the same validation set logic as Table 1.

Model	Method	Category										Average		
		Multi-Hop		Temporal		Open Domain		Single-Hop		Adversarial		Performance*		Task
		F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	Rank
GPT-4o	LoCoMo (Maharana et al., 2024)	28.0	18.5	9.1	5.8	16.5	14.8	**61.6**	**54.2**	52.6	51.1	29.5	22.2	2.8
	ReadAgent (Lee et al., 2024)	14.6	10.0	4.2	3.2	8.8	8.4	12.5	10.3	6.8	6.1	11.7	8.5	5.0
	MemoryBank (Zhong et al., 2024)	6.5	4.7	2.5	2.4	6.4	5.3	8.3	7.1	4.4	3.7	6.0	4.7	6.0
	MemGPT (Packer et al., 2024)	30.4	22.8	17.3	13.2	12.2	11.9	60.2	53.4	35.0	34.3	32.0	25.7	3.2
	A-Mem (Xu et al., 2025)	32.9	23.8	39.4	31.2	17.1	15.8	48.4	43.0	36.4	35.5	36.1	28.4	2.4
	SYNAPSE (Ours)	**39.3**	**29.5**	**55.5**	**50.3**	**29.5**	**23.9**	46.5	38.8	**97.8**	**97.7**	**43.4**	**35.2**	**1.6**
Qwen-1.5b	LoCoMo (Maharana et al., 2024)	9.1	6.6	4.3	4.0	9.9	8.5	11.2	8.7	40.4	40.2	8.5	6.6	4.0
	ReadAgent (Lee et al., 2024)	6.6	4.9	2.6	2.5	5.3	12.2	10.1	7.5	5.4	27.3	6.3	5.3	5.8
	MemoryBank (Zhong et al., 2024)	11.1	8.3	4.5	2.9	8.1	6.2	13.4	11.0	36.8	34.0	10.0	7.5	3.6
	MemGPT (Packer et al., 2024)	10.4	7.6	4.2	3.9	13.4	11.6	9.6	7.3	31.5	28.9	9.1	7.0	4.6
	A-Mem (Xu et al., 2025)	18.2	11.9	24.3	19.7	16.5	**14.3**	23.6	19.2	46.0	43.3	20.4	15.0	2.0
	SYNAPSE (Ours)	**38.1**	**24.6**	**35.5**	**28.6**	**18.1**	11.7	**35.8**	**26.6**	**98.1**	**60.1**	**35.9**	**25.0**	**1.0**
Qwen-3b	LoCoMo (Maharana et al., 2024)	4.6	4.3	3.1	2.7	4.6	6.0	7.0	5.7	17.0	14.8	4.7	4.3	4.8
	ReadAgent (Lee et al., 2024)	2.5	1.8	3.0	3.0	5.6	5.2	3.3	2.5	15.8	14.0	2.9	2.4	5.8
	MemoryBank (Zhong et al., 2024)	3.6	3.4	1.7	2.0	6.6	6.6	4.1	3.3	13.1	10.3	3.5	3.3	6.0
	MemGPT (Packer et al., 2024)	5.1	4.3	2.9	3.0	7.0	7.1	7.3	5.5	14.5	12.4	5.2	4.4	4.6
	A-Mem (Xu et al., 2025)	12.6	9.0	27.6	25.1	7.1	7.3	17.2	13.1	27.9	25.2	16.2	13.0	2.6
	MemoryOS (Li et al., 2025)	21.4	15.0	26.2	22.4	10.2	8.2	23.3	15.4	–	–	22.1	16.2	–
	SYNAPSE (Ours)	**38.8**	**25.1**	**36.2**	**29.6**	**14.7**	**11.5**	**37.8**	**26.1**	**98.9**	**60.5**	**36.6**	**25.4**	**1.0**