# Chronic-Mem: Enhancing Medical Agents Memory via Hierarchical Clinical-Semantic Significance Metabolism for Chronic Disease Management

**Anonymous ACL submission**


## Abstract

As chronic disease management moves toward decentralized home care, medical agents must preserve accurate longitudinal patient histories while filtering large volumes of noisy interaction data. Existing memory systems are largely time-centric: they apply homogeneous decay over past information and therefore struggle to separate transient social details from persistent clinical anchors such as allergies, medications, and chronic conditions. We present Chronic-Mem, a clinical significance-aware memory metabolism framework that shifts memory management from physical time to semantic criticality. Chronic-Mem combines semantic triage ($F_1$), event aggregation ($F_2$), and a tri-state metabolic pipeline ($F_3$) to implement non-homogeneous forgetting dynamics that retain clinically essential evidence while compressing low-value context. Experiments on C-Trap, our adversarial benchmark for chronic-care memory, the public MediLongChat and MedMT-Bench show that Chronic-Mem consistently outperforms existing memory systems in clinical reasoning, long-range recall, and adversarial robustness, while achieving substantial efficiency gains, including more than 5× lower memory overhead and 82.7% context compression.


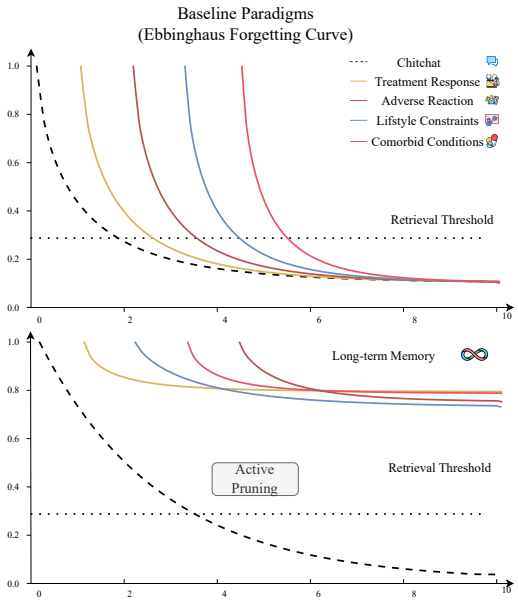


Figure 1: Comparison of memory decay paradigms. Unlike traditional homogeneous decay (top) that inevitably causes critical clinical anchors to fall below the retrieval threshold, our Chronic-Mem (bottom) employs non-homogeneous metabolism to actively prune low-value chitchat while selectively consolidating vital medical evidence into long-term memory.

## 1 Introduction

As Chronic Disease Management (CDM) shifts from hospital-centric to home-based care (Barajas-Nava et al., 2022; Xu et al., 2024; Gu et al., 2024), medical agents with long-term interactive capabilities have become increasingly important for continuous out-of-hospital follow-up (Wang et al., 2024; Xue et al., 2026). These agents are expected not only to answer isolated medical questions, but also to support longitudinal monitoring, symptom interpretation, medication reminders, and risk-sensitive decision support across repeated interactions. In this long-horizon setting, patient history is not merely a stack of dialogue turns, but a structured trajectory of evolving clinical states—a continuous progression where past medical events dynamically shape current diagnostic logic. This progression necessitates the precise retention of critical clinical anchors, such as treatment responses, adverse reactions, lifestyle constraints, and comorbid conditions. However, over months or years of continuous interaction, these transient, high-value medical evidences inevitably become submerged beneath a deluge of daily redundancies, such as emotional expressions or social trivialities (Mama et al., 2025; Dubois et al., 2017). This severe information dilution makes selective, Clinical-Semanic-aware memory management a fundamental necessity for reli-

able chronic-care agents, as illustrated in Figure 1.

To address this challenge, medical agents require dynamic memory lifecycle management—encompassing selective retention, semantic consolidation, and active forgetting—rather than isolated eviction rules (Hu et al., 2025). However, current memory systems critically lack nuanced importance modeling for clinical scenarios. Mainstream approaches—such as retrieval-augmented generation (RAG) (Karpukhin et al., 2020; Lewis et al., 2020; Gao et al., 2025) and context-window extension (Vaswani et al., 2017; Liu et al., 2024)—largely rely on binary or temporally homogeneous retention policies. They either preserve history indiscriminately, triggering information overload (Logan, 2026), or discard historical context based solely on recency, leading to the catastrophic forgetting of critical clinical anchors (Wang et al., 2026). In longitudinal medical settings, both extremes are highly problematic. Unfiltered accumulation obscures crucial diagnostic signals beneath conversational noise, making it harder for the agent to surface relevant evidence at the moment of decision. Conversely, rigid truncation prematurely erases long-standing medical facts (e.g., drug allergies or past interventions) that must remain structurally stable over time. Consequently, as interaction history scales, these models fail to balance completeness with relevance: attention to key evidence degrades (Yu et al., 2024), and forced compression inevitably introduces irreversible information loss (Zhang et al., 2026; Yu et al., 2026). **Ultimately, advancing chronic-care agents requires not simply expanding memory capacity, but establishing a principled mechanism to dynamically decide what must persist, what should be consolidated, and what can be safely forgotten.**

Critically, directly applying cognitive forgetting mechanisms (Murre and Dros, 2015) fails to resolve this dilemma. Existing memory architectures (e.g., MemoryBank (Zhong et al., 2024), FADEMEM (Wei et al., 2026)) overwhelmingly rely on time-centric, homogeneous decay, assuming information importance decreases monotonically. This premise is deeply misaligned with medical reality, where clinical value is dictated by intrinsic semantic significance rather than recency. A historical drug allergy remains absolutely decisive, whereas a recent superficial symptom may offer negligible utility. Thus, uniform temporal decay inevitably forces vital anchors to be overwritten by recent noise, generating fluent yet potentially hazardous recommendations.

To shift from a time-centric to a semantic-centric paradigm, we propose Chronic-Mem, a clinical significance-aware memory metabolism framework. Our core contributions are summarized as follows:

- We propose a clinical significance-aware forgetting mechanism that moves memory management from time-based decay to medically grounded importance. This design helps preserve core clinical anchors even when they appear far back in the interaction history.
- We introduce a multi-level memory architecture that organizes patient information from atoms to events and episodes. With Clinical Significance-aware Compression (CSPC), the framework converts scattered interactions into compact long-term memories that are easier to use for clinical reasoning.
- We build C-Trap, an adversarial benchmark with 100 guideline-based "logic traps" for chronic-care settings. It evaluates whether an agent can still recover the right evidence chain under heavy noise and long-range temporal shifts.
- Experiments on C-Trap, MediLongChat, and MedMT-Bench show that Chronic-Mem consistently improves long-range anchor recall, reaching an A-Recall of approximately 0.79, and strengthens cross-dialogue reasoning with a CDR F1 of approximately 0.392. It also improves robustness to adversarial context interference by up to 81.82%, reduces memory overhead by over 5× compared with MemoryOS, and compresses dialogue context by 82.7%, validating clinically guided hierarchical selective retention for long-horizon medical agents.

## 2 Related Work

Long-term chronic disease management requires medical agents to continuously track patient-specific information across repeated interactions, including treatment response, adverse reactions, lifestyle constraints, or comorbid conditions. However, existing LLM memory mechanisms are not designed to distinguish clinically persistent evidence from low-value conversational noise. Based on how they store, retrieve, and discard information, prior work can be grouped into three paradigms.

**Flat Storage and Passive Retrieval.** The most direct solutions rely on truncation or external re-

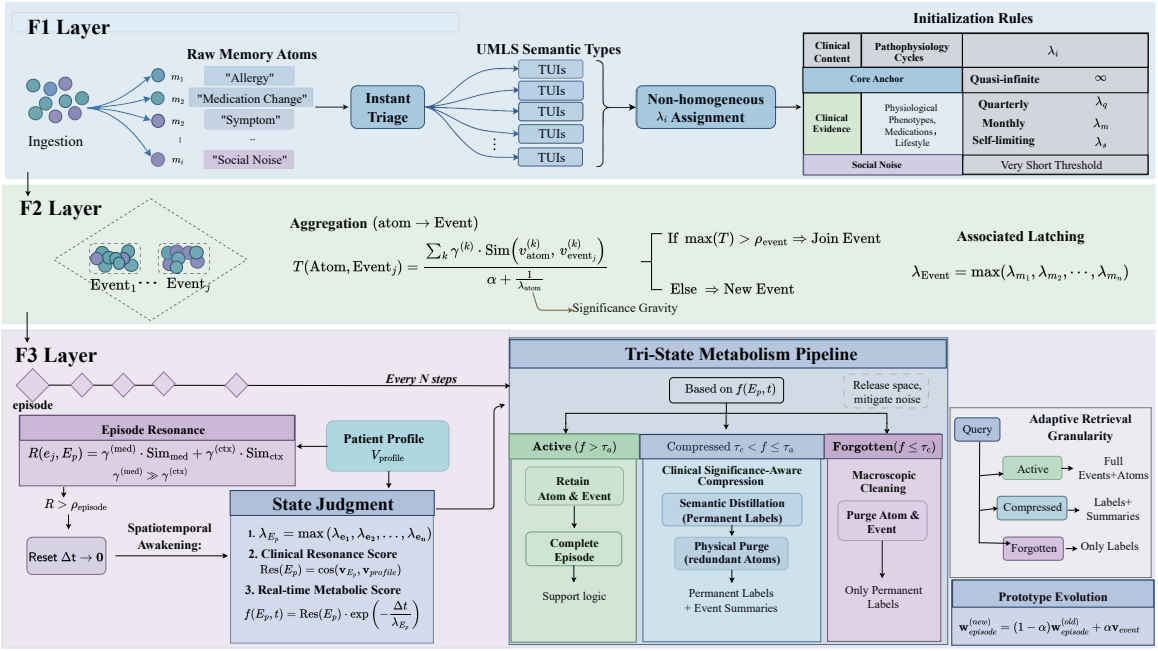


Figure 2: Overview of the Chronic-Mem Framework. The figure presents the end-to-end pipeline of Chronic-Mem across three hierarchical stages. (1) The $F_1$ layer (Semantic Triage & Atomization) decomposes raw dialogues into "Memory Atoms," assigns UMLS (Unified Medical Language System) semantic types, and applies non-homogeneous lifecycle parameters $\lambda$ according to clinical significance (e.g., allergies vs. social noise). (2) The $F_2$ layer (Event Aggregation) clusters related atoms into cohesive "Events" via a "Significance Gravity" mechanism, while an "Associated Latching" rule preserves events according to their most critical atoms. (3) The $F_3$ layer (Episode Metabolism & Tri-State Pipeline) adaptively transitions episodes among "Active," "Compressed," and "Forgotten" states based on clinical resonance and metabolic scores.

trieval. Sliding-window methods keep only the most recent dialogue turns with a strict First-In-First-Out (FIFO) policy, which is efficient but easily discards early patient history that may remain clinically important over long-term follow-up (Vaswani et al., 2017; Vinyals and Le, 2015). Native Retrieval-Augmented Generation (RAG) instead stores dialogue history in an external memory and retrieves records by semantic similarity (Karpukhin et al., 2020; Lewis et al., 2020; Gao et al., 2025). Although effective in open-domain settings, this passive retrieval paradigm is vulnerable in clinical dialogue because lexical overlap does not reliably reflect medical relevance. As a result, clinically decisive facts may be missed, while accumulated daily chatter further lowers retrieval precision.

**Time-Centric Homogeneous Forgetting.** To reduce overload, another line of work introduces active forgetting. MemoryBank draws on the Ebbinghaus forgetting curve to decay memories over time (Murre and Dros, 2015; Zhong et al., 2024); FadeMem uses dual-layer differential forgetting to improve retrieval efficiency (Wei et al., 2026). Despite these design differences, they share a homogeneous decay assumption: memory lifespan is determined mainly by time or frequency. This assumption is reasonable for casual conversation but problematic in healthcare, where old but clinically critical facts, such as drug allergies or chronic conditions, often matter more than recent social details. Without distinguishing clinical anchors from low-value redundancy, such methods risk catastrophic semantic forgetting.

**Structural Consolidation without Medical Priors.** Recent works explore autonomous memory architectures to resolve unstructured storage chaos. MemOS adapts hierarchical storage from operating systems (Li et al., 2025), while A-Mem empowers agents to manage memory fragments via self-reflection (Xu et al., 2025). Despite these structural advancements, these frameworks remain domain-agnostic and lack an awareness of "clinical significance." Consequently, their retrieval efficacy continues to degrade when flooded with months of high-frequency, noisy chronic disease check-ins

(Sharma, 2025; Akbar et al., 2025).Recently, MemoryART leveraged Adaptive Resonance Theory to build a tripartite cognitive architecture for medical dialogues (Dai et al., 2026). However, lacking an inherent assessment of clinical importance, it remains prone to overlooking low-frequency yet critical evidence (e.g., allergy history) amidst a deluge of daily chitchat noise.

In summary, existing memory paradigms lack an intrinsic perception of medical semantic importance. To bridge this gap, our Chronic-Mem shifts the evaluation metric from physical time to clinical significance. By introducing a non-homogeneous memory metabolism mechanism, we fundamentally address the dilemma of high-noise forgetting inherent in long-term medical interactions.

## 3 Model Details

To overcome the inherent limitations of homogeneous time decay in long-term memory systems, we present Chronic-Mem, a clinical significance-aware memory metabolism framework. While we adopt the bottom-up Atom–Event–Episode topology inspired by recent hierarchical episodic memory networks like MemoryART (Dai et al., 2026), existing architectures generally rely on passive self-organizing resonance and lack active, domain-specific forgetting mechanisms. To bridge this gap, Chronic-Mem integrates non-homogeneous forgetting dynamics into the hierarchical network for the first time. This integration transitions memory management from passive clustering to active semantic metabolism, ensuring that the lifecycles of stored patient interactions are governed by clinical criticality rather than chronological order. as illustrated in Figure 2.

**Heterogeneous Decay.** Conventional models often assume that the value of information decays homogeneously over time. We challenge this assumption by proposing a medical semantics-driven heterogeneous decay model. In Chronic-Mem, the survival probability (or retrieval weight) $f(m_i, t)$ of any memory atom node $m_i$ at time $t$ is defined as:

$$f(m_i, t) = \text{Res}(m_i, q) \cdot \exp\left(-\frac{\Delta t_i}{\lambda_i}\right)$$

where $\text{Res}(m_i, q)$ denotes the resonance score between the memory $m_i$ and the current query $q$, and $\Delta t_i = t - t_{\text{occurrence}}$ represents the time elapsed since the memory's generation. The Significance Coefficient $\lambda_i$ determines the rate of memory metabolism and is initialized via strict clinical hard-rules.

**Hierarchical Semantic Metabolism Architecture.** To extract clinically useful evidence from noisy long-term dialogue, Chronic-Mem uses a three-stage architecture: $F_1$, $F_2$, and $F_3$.

**$F_1$ Layer: Semantic Triage and Lifecycle Initialization.** When an atomic memory $m_i$ enters the system, $F_1$ first maps its entities to UMLS Semantic Types (TUIs) and initializes its significance coefficient $\lambda_i$. Instead of assigning the same decay rule to all memories, Chronic-Mem sets $\lambda_i$ according to clinical decision value:

- Core Safety Anchors (e.g., allergies): Assigned a quasi-infinite lifecycle to form a stable safety baseline.
- Critical Clinical Information (e.g., medication changes, physiological phenotypes, acute symptoms): Assigned lifecycles that match their typical clinical follow-up or pathological duration.
- Low-Entropy Social Noise: Assigned a minimal survival threshold so redundant content can fade quickly.

**$F_2$ Layer: Event Aggregation and Associated Latching.** Clinically related evidence is often scattered across multiple trivial-looking utterances. The $F_2$ layer therefore groups atoms into coherent clinical Events through "Significance Gravity" and preserves important signals through Associated Latching.

**Aggregation (Atom $\rightarrow$ Event).** The driving force $T$ for aggregation is defined as:

$$T(\text{Atom}, \text{Event}_j) = \frac{\sum_k \gamma^{(k)} \cdot \text{Sim}(v^{(k)}_{\text{atom}}, v^{(k)}_{\text{event}_j})}{\alpha + \frac{1}{\lambda_{\text{atom}}}}$$

The numerator measures similarity, while the denominator acts as "Significance Gravity." Atoms with larger $\lambda$ exert stronger pull and are more likely to absorb surrounding details into the same event. If $\max(T) > \rho_{\text{event}}$, the atom is merged into an existing event; otherwise, a new event is created. Crucially, the threshold $\rho_{\text{event}}$ serves as a tunable semantic boundary that controls the granularity of event aggregation, ensuring that distinct clinical episodes are not over-generalized or prematurely conflated.

**Associated Latching (Max-Pooling).** To preserve the full evidence chain, each event inherits the lifecycle of its most durable atom:

$$\lambda_{\text{Event}} = \max(\lambda_{m_1}, \lambda_{m_2}, \ldots, \lambda_{m_n})$$

$F_3$ **Layer: Episode Resonance and Pattern Elevation.** To alleviate exponentially growing contextual redundancy in long-term medical management, the $F_3$ layer introduces a Tri-State Metabolism mechanism that dynamically elevates chronologically discrete, low-order Events into high-order Episodes. Conceptually, an Episode represents a longitudinal disease trajectory line aligned across multiple clinical encounters via core medical semantics.

**Episode Resonance (Event → Episode).** To capture the trajectory of disease evolution across extended temporal spans, the network processes input event sequences chronologically by computing the cross-temporal semantic resonance $R(e_j, E_p)$ between the current event $e_j$ and an existing episode $E_p$:

$$R(e_j, E_p) = \gamma^{(med)} \cdot \text{Sim}_{med}(v_{e_j}^{(med)}, w_{E_p}^{(med)}) + \gamma^{(ctx)} \cdot \text{Sim}_{ctx}(v_{e_j}^{(ctx)}, w_{E_p}^{(ctx)})$$

where $v_{e_j}$ and $w_{E_p}$ denote the feature encoding of the event and the prototype vector of the episode, respectively. We explicitly enforce the constraint $\gamma^{(med)} \gg \gamma^{(ctx)}$ to ensure that core clinical words dominate the resonance mapping, thereby bypassing the naive limitations of relying solely on physical temporal intervals.

If the maximum resonance score satisfies $\max_p R(e_j, E_p) > \rho_{episode}$, where $\rho_{episode}$ is a tunable semantic boundary, $e_j$ is aggregated into $E_p$. Subsequently, the episode prototypes $(w_{med}, w_{ctx})$ are dynamically updated via an Exponential Moving Average (EMA) to model incremental semantic drift over time. Concurrently, this integration triggers a spatiotemporal awakening mechanism that resets the relative elapsed time ($\Delta t \to 0$), thereby prioritizing the episode's survival in downstream metabolic evaluations. Otherwise, $e_j$ serves as a seed node to independently initialize a new Episode.

**State Judgment.** The baseline lifespan of each episode $E_p$ is determined by its constituent events using max-anchoring:

$$\lambda_{E_p} = \max(\lambda_{e_1}, \lambda_{e_2}, \ldots, \lambda_{e_n}), \quad e_k \in E_p$$

The system also maintains a global patient profile vector $\mathbf{v}_{profile}$ that captures stable clinical attributes such as age, sex, and diagnosed conditions. Every $N$ turns, background metabolism computes the resonance between each episode and this profile:

$$\text{Res}(E_p) = \cos(\mathbf{v}_{E_p}, \mathbf{v}_{profile})$$

Combined with time decay, the real-time metabolic score is:

$$f(E_j, t) = \text{Res}(E_p) \cdot \exp\left(-\frac{\Delta t}{\lambda_{E_p}}\right)$$

where $\Delta t$ is the time since the episode was last awakened. This means that even older episodes can remain active if they still strongly match the patient's long-term clinical condition.

**The Tri-State Metabolic Pipeline.** Based on the real-time score $f$, Chronic-Mem follows a "semantic distillation first, physical purge second" strategy.

- Active Phase ($f > \tau_a$): The episode remains fully stored, including both Atoms and Events, so the agent can access fine-grained evidence for recent clinical reasoning.
- Compressed Phase ($\tau_c < f \leq \tau_a$): The system invokes CSPC to retain key pathological features as Permanent Labels and deletes redundant atoms that no longer add value.
- Forgotten Phase ($f \leq \tau_c$): The system removes the underlying atoms and aggregated vectors, keeping only Permanent Labels as compact long-term indices.

**Adaptive Retrieval Granularity.** The state of an episode directly determines what is returned at query time. An Active hit returns "Events + fine-grained Atom details," a Compressed hit returns "Permanent Labels + Event summaries," and a Forgotten hit returns only "Permanent Labels."

**Episode Evolution.** To ensure the episode's centroid reflects the patient's latest disease progression, the system updates the prototype using a moving average:

$$\mathbf{w}_{episode}^{(new)} = (1 - \alpha)\mathbf{w}_{episode}^{(old)} + \alpha \mathbf{v}_{event}$$

This design ensures that even during months-long chronic care, the system can precisely locate relevant historical disease segments via "semantic resonance," leveraging Permanent Labels to achieve a paradigm shift from volatile physical storage to solidified permanent semantic knowledge.

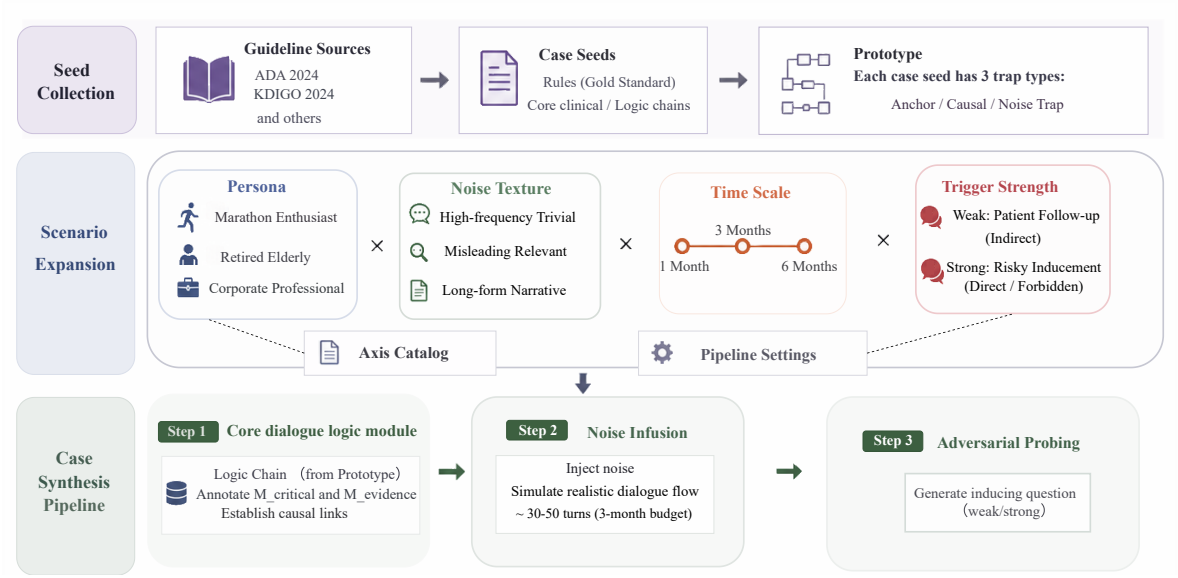


Figure 3: Construction pipeline of the C-Trap benchmark. Starting from guideline-grounded seed cases, the pipeline expands scenarios through persona and noise variation, stretches dialogue trajectories across chronic-care timelines, implants adversarial probing targets, and produces structured annotations for evaluating long-horizon clinical memory.

## 4 Benchmark C-Trap

Existing benchmarks like MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019) predominantly evaluate single-turn or short-context reasoning. However, in real longitudinal follow-ups, critical signals are often submerged in conversational noise, making isolated symptom retrieval inadequate without linking to prior interventions. To bridge this gap, we introduce **C-Trap** (Clinical-Semantic Temporal Resilience Anchor Probe), a guideline-based benchmark featuring long-horizon logical traps that challenge models to recover coherent clinical evidence chains under high-noise conditions.

**Construction Pipeline.** We establish the C-Trap benchmark via a seed-guided scenario evolution paradigm (Figure 3). First, we extract core causal chains from authoritative guidelines to serve as grounded clinical seeds (Whelton et al., 2018; American Diabetes Association, 2025; Levin et al., 2024). To simulate the low signal-to-noise ratio of longitudinal care, we introduce adversarial noise by crossing distinct patient personas with high-frequency social distractors, spanning 1-to-6 month dialogue trajectories. Finally, we implant adversarial probes at the session tail and leverage LLMs to annotate the decision logic, structuring the dataset with atomic annotations of critical events and supporting evidence.

**Evaluation Metrics.** To comprehensively quantify the agent's reliability in long-context diagnostics, we establish a three-dimensional evaluation framework.

**Safety.** This dimension measures the model's ability to maintain critical medical logic under extreme interference. We evaluate A-Recall (Anchor Recall), which quantifies the retention of "lifeline" medical history (e.g., absolute contraindications), defined as the ratio of successfully retrieved core anchors to the total number of embedded anchors.

**Efficiency.** To validate Chronic-Mem's intelligent metabolism mechanism, we define Metabolic Efficiency (ME). This metric evaluates the storage cost-effectiveness achieved through intelligent forgetting. It reflects the clinical decision-making value carried per unit of compressed space after the physical purge of redundant information:

$$\text{ME} = \frac{\text{Count}(M_{\text{anchor}})}{\text{Storage}_{\text{compressed}} / \text{Storage}_{\text{raw}}}$$

**System Overhead.** We also monitor offline metabolic latency (ML) and storage footprint (SF). ML measures the background cost of vector recomputation, atom pruning, and state switching. SF reflects the resource efficiency of long-term management on edge devices, serving as a key indicator of engineering feasibility.

**Benchmark Validity.** On a stratified 20-case validation subset, oracle anchor injection achieves per-

| Method | A-Recall ↑ | ME ↑ | ML ↓ | SF ↓ |
|---|---|---|---|---|
| Native-RAG | 0.14 | 0.0017 | 6090.15 | 0.276 |
| A-Mem | 0.09 | 0.0111 | 2908.61 | 0.101 |
| MemoryBank | 0.47 | 0.0573 | **134.57** | **0.052** |
| MemoryART | 0.24 | 0.0383 | 5162.19 | 1.089 |
| MemoryOS | 0.61 | 0.0014 | 52033.04 | 4.529 |
| **Chronic-Mem** | **0.778** | **0.1734** | 15734.78 | 0.187 |

Table 1: Averaged performance across the DeepSeek-V4-Pro and GPT-4o-mini backbones. Results are arithmetic means over two representative LLM settings. Best results are in bold.

fect A-Recall (1.00), while restricting access to only the final three turns reduces A-Recall to 0.03, demonstrating that the benchmark is both solvable and highly challenging. All gold decisions are fully guideline-consistent, and an independent LLM auditor rated 100% of gold labels as clinically aligned. And we use manual sampling testing to ensure that the large model does not exhibit hallucinations. See Appendix A for details.

# 5 Experiments

## 5.1 Experiment Settings

**Datasets.** We evaluate Chronic-Mem on two datasets: (1) C-Trap, our proposed adversarial diagnostic benchmark designed to stress-test long-context clinical reasoning and evidence recovery under massive noise; and (2) MediLongChat (Hu et al., 2026), a synthetic long-context healthcare dialogue dataset that simulates realistic lifelong patient–physician interactions for evaluating long-term medical memory and reasoning. (3) To further demonstrate the effectiveness of Chronic-Mem in emerging long-context medical interaction settings, we additionally evaluate our framework on the recently proposed MedMT-Bench (Yang et al., 2026). Detailed results and analysis are provided in Appendix B.

**Evaluation Metrics.** For C-Trap, we assess retrieval quality and system efficiency using A-Recall (Anchor Recall), ME (Metabolic Efficiency), ML (Memory Latency, in seconds), and SF (Storage Footprint, in MB). For MediLongChat, we compute the micro-averaged BLEU and F1 scores for both In-dialogue Reasoning (IDR) and Cross-dialogue Reasoning (CDR) to assess text generation fidelity across long-horizon dialogues. Additionally, we report the Memory/Raw Ratio to directly quantify the context compression efficiency of the memory system against dense historical transcripts.

**Compared Methods.** We benchmark Chronic-Mem against the vanilla backbone without memory, a Native-RAG baseline (Lewis et al., 2020), and four state-of-the-art long-term memory frameworks: A-Mem (Xu et al., 2025), MemoryBank (Zhong et al., 2024), MemoryOS (Kang et al., 2025), and MemoryART (Dai et al., 2026).

## 5.2 Results on C-Trap

Table 1 summarizes the averaged C-Trap performance. We compare six memory methods under our benchmark, which primarily evaluates whether long-range critical clinical anchors can be recalled at probe time, rather than general conversational fluency.

Native RAG and A-Mem show the weakest long-range recall, with A-Recall values of approximately 0.14 and 0.09, respectively. Native RAG is constrained by flat vector retrieval and therefore struggles to recover sparse anchors across distant visits. A-Mem is more efficient in storage and updating, using about 0.10 MB of storage footprint and 2908 ms of memory latency, but its note-centric organization still leaves long-range anchor recall as the main bottleneck. MemoryBank provides a stronger efficiency–recall trade-off: its A-Recall reaches approximately 0.47, while its memory latency and storage footprint remain low at about 134 ms and 0.05 MB. Nevertheless, its Ebbinghaus-style time/frequency decay can prematurely discard clinically relevant but temporally distant facts, leaving it substantially below Chronic-Mem. MemoryART improves over RAG with an A-Recall of about 0.24, yet its larger storage footprint of roughly 1.09 MB yields only limited metabolic efficiency (ME = 0.038), suggesting that episode-level retrieval alone is insufficient without medically grounded selective retention. MemoryOS achieves the strongest baseline recall, with A-Recall around 0.61, but this comes from LLM-heavy ingest and cross-case accumulation, resulting in a much larger storage footprint of about 4.53 MB, memory latency of about 52033 ms, and ME of only 0.0014. By contrast, Chronic-Mem reaches an A-Recall of approximately 0.78 with ME around 0.17 and a compact storage footprint of about 0.19 MB, indicating that its hierarchical metabolic design better fits long-horizon chronic-care traps.

| Method | IDR BLEU ↑ | IDR F1 ↑ | CDR BLEU ↑ | CDR F1 ↑ | Memory / Raw Ratio ↓ |
|---|---|---|---|---|---|
| **Chronic-Mem (Ours)** | **0.176** | **0.485** | **0.058** | **0.392** | **0.033** |
| MemoryBank | 0.113 | 0.367 | 0.025 | 0.245 | 0.043 |
| MemoryOS | 0.145 | 0.448 | 0.048 | 0.366 | 0.171 |
| A-Mem | 0.115 | 0.405 | 0.032 | 0.275 | 0.044 |
| Vanilla (w/o Memory) | 0.126 | 0.437 | 0.040 | 0.296 | N/A |

Table 2: Main experimental results on the MediLongChat benchmark. IDR and CDR denote In-dialogue Reasoning and Cross-dialogue Reasoning, respectively. The best performance is highlighted in **bold**, and the second best is underlined. N/A represents Not Applicable.

| Variant | A-Recall (↑) | Δ | ME (↑) | Δ | SF (MB) (↓) | Δ | ML (ms) |
|---|---|---|---|---|---|---|---|
| A: Full Model | 0.86 | - | 0.1812 | - | 0.252 | - | 1805.97 |
| B: w/o $F_1$ (Triage) | 0.79 | -8.1% | 0.1749 | -3.5% | 0.110 | -56.3% | 1783.06 |
| C: w/o $F_2$ (Aggr.) | 0.60 | -30.2% | 0.0877 | -51.6% | 0.530 | +110.3% | 292.92 |
| D: w/o $F_3$ (Pipe) | 0.435 | -49.4% | 0.0098 | -94.6% | 0.275 | +9.1% | 1099.02 |

Table 3: Ablation study results on the core modules. Δ indicates the relative change compared to the full model.

### 5.3 Results on MediLongChat

Table 2 shows that Chronic-Mem achieves stronger cross-dialogue reasoning (CDR) while maintaining high memory efficiency in long-horizon settings. Its CDR F1 reaches 0.392, compared with 0.366 for MemoryOS and 0.296 for the Vanilla baseline, indicating a stronger ability to propagate historical clinical evidence across sessions. The Vanilla baseline lacks an explicit long-term memory mechanism, while MemoryBank and A-Mem rely on general conversational abstractions that provide compression but do not explicitly filter information by medical importance. Chronic-Mem also achieves these reasoning gains with a Memory/Raw Ratio of only 0.033; by comparison, MemoryOS uses a ratio of 0.171, more than 5× higher. These results suggest that long-horizon clinical performance depends less on maximizing raw content retention and more on selectively preserving evidence with long-term diagnostic value.

### 5.4 Ablation Study

We conduct an ablation study with the GPT-4o-Pro backbone (Table 3). Removing any $F$-layer module substantially degrades performance, confirming the synergy of our hierarchical design.

**Core Module Analysis. The $F_3$ Metabolic Pipeline** is the main performance driver: removing it reduces A-Recall by 49.4% and drives Metabolic Efficiency (ME) close to zero, showing the importance of proactive forgetting and label solidification under long-range noise. **The $F_2$ Event Aggregation Layer** balances logic and storage. Disabling $F_2$ increases Storage Footprint (SF) by 111.5% and reduces precision by 30.2%, indicating that clustering atoms into structured events is essential for efficient semantic correlation. **The $F_1$ Semantic Triage Layer** serves as the evidence gatekeeper. Removing $F_1$ lowers A-Recall by 8.1%; although SF decreases without knowledge mapping, the precision loss confirms the need for guideline-based heterogeneous decay to protect clinical anchors.

## 6 Conclusion

We propose Chronic-Mem, a clinical significance-aware memory framework for long-horizon chronic disease management that shifts memory control from time-based decay to medically grounded importance, preserving core clinical anchors even when they appear far back in interaction history. It introduces a multi-level memory architecture that organizes patient information from atoms to events and episodes, and uses CSPC to convert fragmented interactions into compact, reusable long-term memories for clinical reasoning. To evaluate robustness in realistic settings, we construct C-Trap, an adversarial benchmark with 100 guideline-based logic traps to test whether agents can recover correct evidence chains under heavy noise and long-range temporal shifts. Experiments on C-Trap, MediLongChat, and MedMT-Bench show that Chronic-Mem consistently improves long-range A-recall and cross-dialogue reasoning, while enhancing robustness to adversarial interference, and achieving strong dialogue compression, demonstrating the effectiveness of clinically guided hierarchical selective retention for long-horizon medical agents.

## Limitations

Chronic-Mem still partly relies on expert-defined salience priors and guideline mappings, which may need adaptation across specialties, institutions, and patient populations. Our evaluation is limited to text-based chronic-care dialogues and benchmarked memory settings; deployment will require validation with multimodal records, real clinical workflows, and prospective safety monitoring. The framework improves long-horizon evidence retention and compression, but does not replace clinician oversight in high-stakes care.

## Ethics Statement

This work investigates memory management frameworks for long-horizon medical dialogue agents and is intended strictly for research evaluation rather than direct clinical deployment. Chronic-Mem is a supportive information-management tool and must not be used as an autonomous clinical diagnostic or treatment system; any real-world application requires rigorous clinician oversight, prospective safety monitoring, and alignment with institutional privacy, security, and informed consent standards.

Regarding benchmark construction, our evaluation protocols leverage stylized, guideline-grounded synthetic chronic-care scenarios designed specifically to exclude any personally identifiable information (PII) or real patient records, thereby mitigating privacy risks. More broadly, we acknowledge that deploying long-term memory architectures in healthcare agents demands strict adherence to data minimization, definitive retention schedules, stringent access control, and comprehensive audit logs. These governance measures are critical to prevent privacy leakage, eliminate algorithmic biases toward vulnerable patient demographics, and mitigate the risks of unsafe clinical reliance on outdated medical information.

## References


Nur Arifin Akbar, Rahool Dembani, Biagio Lenzitti, and Domenico Tegolo. 2025. RAG-driven memory architectures in conversational LLMs: a literature review with insights into emerging agriculture data sharing. *IEEE Access*.

American Diabetes Association. 2025. 7. diabetes technology: standards of care in diabetes—2025. *Diabetes Care*, 48(Supplement_1):S146–S166.

Leticia A Barajas-Nava, Juan Garduño-Espinosa, Juan M Mireles Dorantes, Raúl Medina-Campos, and M Carmen García-Peña. 2022. Models of comprehensive care for older persons with chronic diseases: a systematic review with a focus on effectiveness. *BMJ Open*, 12(8):e059606.

Renke Dai, Hebin Hu, Jiahui Zhang, Yilin Kang, and Ah-Hwee Tan. 2026. Memoryart: Enhancing LLMs via multi-memory models with adaptive resonance theory for healthcare agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20676–20683.

Sebastien Dubois, Nathanael Romano, David C Kale, Nigam Shah, and Kenneth Jung. 2017. Effective representations of clinical notes. *arXiv preprint arXiv:1705.07025*.

Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. 2025. Synergizing RAG and reasoning: A systematic review. *arXiv preprint arXiv:2504.15909*.

Shuyan Gu, Cangcang Jia, Fangfang Shen, Xiaoyong Wang, Xiaoling Wang, and Hai Gu. 2024. The effects of community-based home health care on the physical and mental health of older adults with chronic diseases. *Quality of Life Research*, 33(3):691–703.

Hebin Hu, Renke Dai, Ah-Hwee Tan, and Yilin Kang. 2026. Synthesis and evaluation of long-term history-aware medical dialogue. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*. To appear.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25972–25981.

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.

Adeera Levin, Sofia B Ahmed, Juan Jesus Carrero, Bethany Foster, Anna Francis, Rasheeda K Hall, Will G Herrington, Guy Hill, Lesley A Inker, Rümeyza Kazancıoğlu, and 1 others. 2024. Executive summary of the kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease: known knowns and known unknowns. *Kidney International*, 105(4):684–701.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, and 1 others. 2025. MemOS: A memory OS for AI systems. *arXiv preprint arXiv:2507.03724*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Joe Logan. 2026. Continuum memory architectures for long-horizon LLM agents. *arXiv preprint arXiv:2601.09913*.

Eden Mama, Liel Sheri, Yehudit Aperstein, and Alexander Apartsin. 2025. From fuzzy speech to medical insight: Benchmarking LLMs on noisy patient narratives. *arXiv preprint arXiv:2509.11803*.

Jaap MJ Murre and Joeri Dros. 2015. Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, 10(7):e0120644.

Chaitanya Sharma. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Bijun Wang, Onur Asan, and Yiqi Zhang. 2024. Shaping the future of chronic disease management: Insights into patient needs for AI-based homecare systems. *International Journal of Medical Informatics*, 181:105301.

Zhenting Wang, Huancheng Chen, Jiayun Wang, and Wei Wei. 2026. Memex (RL): Scaling long-horizon LLM agents via indexed experience memory. *arXiv preprint arXiv:2603.04257*.

Lei Wei, Xiao Peng, Xu Dong, Niantao Xie, and Bin Wang. 2026. Fademem: Biologically-inspired forgetting for efficient agent memory. *arXiv preprint arXiv:2601.18642*.

Paul K Whelton, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, Samuel Gidding, Kenneth A Jamerson, Daniel W Jones, and 1 others. 2018. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248.

Tingke Xu, Zishuo Huang, Bingzhen Li, Haojie Jin, Jiayun Zhang, Huiting Yang, Yucheng Huang, Xiangyang Zhang, and Chun Chen. 2024. Association between home and community-based services utilization and self-rated health among chinese older adults with chronic diseases: evidence from the 2018 china health and retirement longitudinal study. *BMC Public Health*, 24(1):117.

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

Zhikai Xue, Tianqianjin Lin, Pengwei Yan, Ruichun Wang, Yuxin Liu, Zhuoren Jiang, and Xiaozhong Liu. 2026. VitalDiagnosis: AI-driven ecosystem for 24/7 vital monitoring and chronic disease management. *arXiv preprint arXiv:2601.15798*.

Lin Yang, Yuancheng Yang, Xu Wang, Changkun Liu, and Haihua Yang. 2026. Medmt-bench: Can LLMs memorize and understand long multi-turn conversations in medical scenarios? *arXiv preprint arXiv:2603.23519*.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024. In defense of RAG in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. 2026. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. *arXiv preprint arXiv:2601.01885*.

Kehao Zhang, Shangtong Gui, Sheng Yang, Wei Chen, and Yang Feng. 2026. Learning to remember: End-to-end training of memory agents for long-context reasoning. *arXiv preprint arXiv:2602.18493*.

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

# A C-Trap Benchmark Audits and Validity Checks

This appendix supplements Section 5 (The C-Trap Benchmark) with dataset-level audits and validity checks that are summarized in the main text. We report (i) full-corpus statistics over all 100 cases in the CTrap_n100 release, including guideline-seed coverage, surface-axis marginals, and long-horizon structural stress measures; and (ii) targeted sanity checks on a stratified 20-case subset (four cases per seed), covering oracle versus no-early-context A-Recall and gold-label quality assessed by keyword rules and an independent LLM rubric (blind to model outputs). Agent-level scores (A-Recall, ME, ML, SF) and system comparisons appear in the main evaluation tables; material here establishes that the benchmark instances are guideline-grounded, temporally stretched, and non-trivial for long-range memory.

Table A.1: Guideline seed coverage and trap-type distribution in C-Trap (`CTrap_n100`, $n$=100).

| Seed ID | Guideline | Trap | $n$ |
|---|---|---|---|
| R-HTN-01 | ACC/AHA 2017 | A | 20 |
| R-DM-02 | ADA 2024 | B | 20 |
| R-CKD-03 | KDIGO 2024 | A | 20 |
| R-ASTH-04 | GINA 2024 | C | 20 |
| R-HTN-05 | ACC/AHA 2017 | A | 20 |
| Trap-type marginal | | | A:60, B:20, C:20 |

Table A.2: Marginal distribution of surface-variation axes ($n$=100).

| Axis | Level | Count | Share |
|---|---|---|---|
| Persona | A | 30 | 30% |
| | B | 33 | 33% |
| | C | 37 | 37% |
| Noise texture | T1 | 37 | 37% |
| | T2 | 31 | 31% |
| | T3 | 32 | 32% |
| Temporal stretch | 3 months | 100 | 100% |
| Trigger intensity | strong | 100 | 100% |

Table A.3: Structural statistics confirming long-horizon, low-anchor-density stress ($n$=100).

| Statistic | Mean | Median | Min | Max |
|---|---|---|---|---|
| Dialogue turns | 45.8 | 45 | 35 | 58 |
| Anchor–probe distance | 40.0 | 41 | 16 | 55 |
| Anchors per case | 1.42 | 1 | 1 | 2 |
| Noise turns | 39.6 | 39.5 | 30 | 50 |

Table A.4: Sanity check on a stratified 20-case subset (4 per seed).

| Setting | Memory for scoring | Mean A-Recall |
|---|---|---|
| Oracle | Injected anchor summaries | 1.00 |
| No early context | Last 3 turns only | 0.03 |

Table A.5: Gold-label quality audit ($n$=20 stratified).

| Audit | Metric | Result |
|---|---|---|
| Keyword vs. seed | Consistent direction | 20/20 (100%) |
| LLM rubric | Guideline alignment $\geq$4/5 | 20/20 (100%) |
| LLM rubric | Mean alignment / salience | 5.0 / 5.0 |

## B Extended Experiments and Scalability Analysis on MedMT-Bench

This appendix provides additional experimental details and scalability analyses for MedMT-Bench. It is intended to supplement the main results without introducing claims that are necessary for understanding the core method or primary conclusions.

### B.1 Dataset

MedMT-Bench (Yang et al., 2026) is a challenging public benchmark for quantifying LLMs' instruction-following and comprehension capabilities in long, multi-turn medical conversations.

### B.2 Evaluation Metrics

For MedMT-Bench, following its standard protocol, we use an LLM-based evaluator for binary (Yes/No) semantic assessments against annotated test points and report both overall and fine-grained pass rates.

### B.3 Results on MedMT-Bench

Table B.1 and the fine-grained analyses in Figures B.1 and B.2 present the performance of Chronic-Mem in real-world long-form medical dialogue scenarios. Overall, the results provide strong evidence that our semantic-centric memory paradigm improves complex clinical decision-making under long-horizon interaction settings.

**Performance and Compression Efficiency.** As shown in Table B.1, Chronic-Mem achieves a pass rate of 62.0%, substantially outperforming strong baselines such as MemoryART (Dai et al., 2026), which reaches 43.0%. This advantage is accompanied by a striking compression benefit: while MemoryART (Dai et al., 2026) and MemoryBank (Zhong et al., 2024) often saturate the context with nearly 80,000 characters of historical dialogue, Chronic-Mem keeps the average input length to only 13,857 characters, corresponding to an approximate 82.7% reduction. Despite this minimalist input, Chronic-Mem produces the longest responses on average (2,073 characters) while preserving informativeness. This result suggests that the tri-state metabolism architecture enables the model to rely on distilled high-order semantic labels rather than brute-force context accumulation.

**Scenario Adaptability and Multidimensional Generalization.** Chronic-Mem also shows strong stability across diverse medical scenarios. As indicated by the fine-grained analyses in Figures B.1 and B.2, it substantially outperforms the baselines in the resisting-context-interference dimension, reaching 81.82%. We attribute this robustness to the long-term preservation of critical clinical landmarks under heterogeneous forgetting. More broadly, Chronic-Mem achieves strong performance across a range of settings, from diagnosis (50.0%) and rehabilitation (54.2%), which require longer causal chains, to high-frequency consultation scenarios (68.9%). These results suggest that the framework generalizes across both complex reasoning and routine follow-up settings, while maintaining stable therapeutic consistency over long-term chronic care.

**Superior Robustness in Adversarial Scenarios.** As shown in Figure B.3, Chronic-Mem demonstrates clear robustness advantages under adversarial settings. Under contradictory dialogue history, the vanilla backbone collapses to a 0% pass rate, while conventional memory frameworks remain limited at around 40.0%. In contrast, Chronic-Mem raises the pass rate to 80.0% by tracing the correct evidence chain more precisely. The same trend appears in high-pressure settings such as comorbidity (87.5%) and multi-person interweaving (85.7%), where Chronic-Mem achieves more than double the performance of the baselines. This finding indicates that the $F_1$ triage layer effectively filters irrelevant symptomatic noise, allowing the model to preserve medically relevant logic even under severe interference.

| Method | n | Pass | Pass Rate ↑ | Latency (ms) ↓ | Input Chars | Output Chars |
|---|---|---|---|---|---|---|
| **Chronic-Mem (Ours)** | 100 | 62 | **62%** | 62571.95 | 13857.89 | 2073.33 |
| MemoryART | 100 | 43 | 43% | 35117.04 | 80069.00 | 2818.19 |
| MemoryBank | 100 | 41 | 41% | 36502.62 | 80069.00 | 2944.26 |
| Vanilla (w/o Memory) | 100 | 37 | 37% | 39643.63 | 5176.58 | 3368.30 |

Table B.1: Performance on the MedMT-Bench. Quantitative results across response quality, context efficiency, and system latency. $n$ denotes the number of test instances. Best results are highlighted in bold.

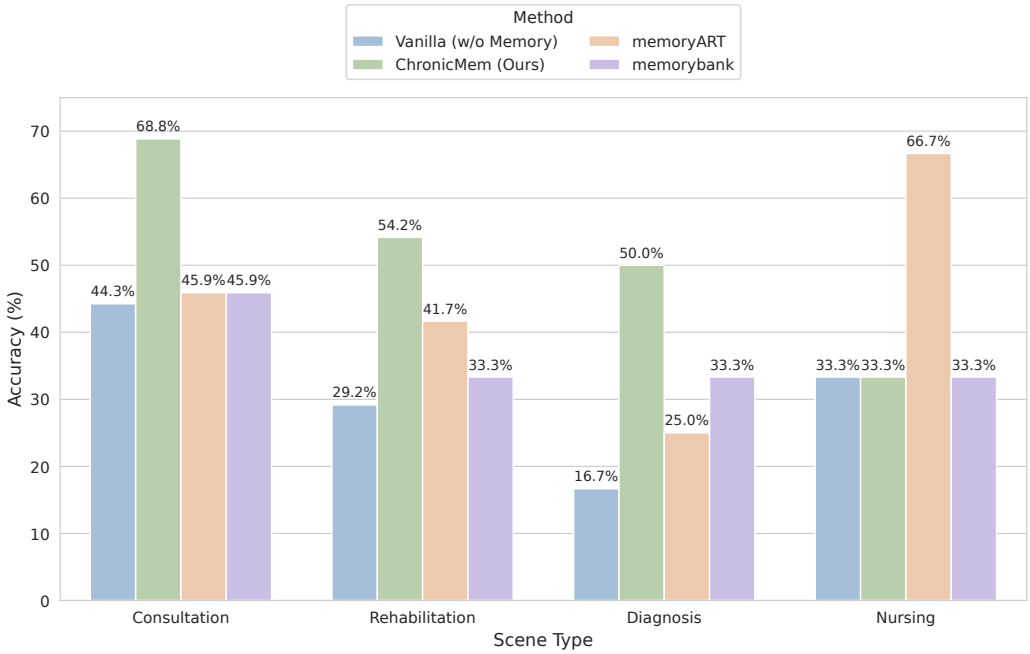


Figure B.1: Performance across diverse clinical scene types on the MedMT-Bench.

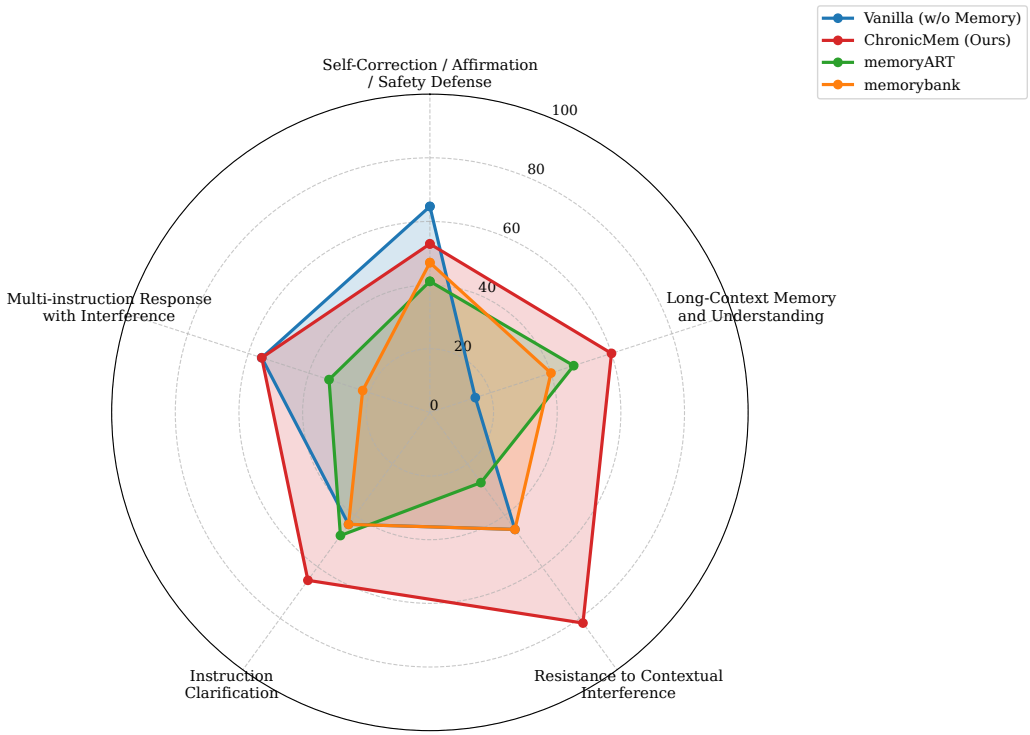


Figure B.2: Instruction-following robustness on the MedMT-Bench.

Figure B.3: Robustness under adversarial sub-instruction types on the MedMT-Bench.