EVALUATING AND UNDERSTANDING AGENT MEMORY SYSTEMS FOR CHRONIC CARE

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory is essential for large language model (LLM) agents in chronic care, where patient histories span years of visits across multiple departments. Many memory systems have been proposed for LLM agents, yet existing benchmarks, drawn from general-domain dialogues or medical records over limited years of visits and departments, cannot reveal how they behave or why they fail in chronic care. To fill this gap, we introduce MEME-Bench, a clinically grounded benchmark of patient histories synthesized from real chronic-care records, with records spanning over a decade, patients up to 1,718 visits, and 1,646 questions on eight capabilities asked as the history accumulates. Leveraging MEME-Bench, we evaluate 11 memory systems and a long-context baseline across seven backbones and uncover three critical findings: (1) no memory system dominates across capabilities, and none outperforms the long-context baseline on every axis; (2) longitudinal trajectory tracking and cross-department integration remain poor for all systems; and (3) tracing the required facts through the memory pipeline reveals retrieval as the most common source of failure, rooted in evidence scattered across departments, recurring measurements and precise medical terminology. Controlled ablations further show that combining agentic retrieval with date-filtered search improves both of the hardest axes, whereas prompting cannot replace these components. MEME-Bench is live on Hugging Face.

1 INTRODUCTION

Chronic diseases account for roughly 7 out of 10 deaths in the United States and over 40 million deaths globally each year (Organization et al., 2018). Managing these conditions places a substantial burden on clinicians, who must draw on ever-growing patient histories to analyze vital trends, adjust medications, and monitor treatment responses over years of follow-up visits (Porter et al., 2023). Large language model (LLM) agents have emerged as a promising approach to alleviate this burden through automated patient monitoring and clinical decision support (Li et al., 2024). For an agent to safely fill this role, it must maintain an accurate understanding of the evolving patient status over time, such as recalling adverse drug reactions, recognizing disease progressions and tracking medication dosage changes. Reliable memory mechanisms are therefore essential for chronic-care agents.

Recent work has made substantial progress in developing memory systems that help LLM agents retain and use information across long interaction histories, ranging from retrieval-augmented generation (Lewis et al., 2020) and knowledge graphs (Gutiérrez et al., 2025) to summarization pipelines (Fang et al., 2026) and agent-controlled memory (Shu et al., 2026). Despite promising results on general-domain benchmarks (Maharana et al., 2024), how these systems behave under realistic chronic-care settings remains largely underexplored.

Existing benchmarks leave this question open for two reasons. First, performance on general-domain benchmarks may not transfer to chronic care, which places unique requirements on the abilities these benchmarks assess. Typically, the general-domain memory benchmarks evaluate capabilities such as evidence recall, temporal reasoning, knowledge update, and abstention (Wu et al., 2024; Maharana et al., 2024; Hu et al., 2026; Tavakoli et al., 2026). However, these assessments do not reflect the complexity of real-world chronic care (Skou et al., 2022), where the agent must recover precise evidence buried within repetitive follow-up visit notes, reason over full trajectories of vitals

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

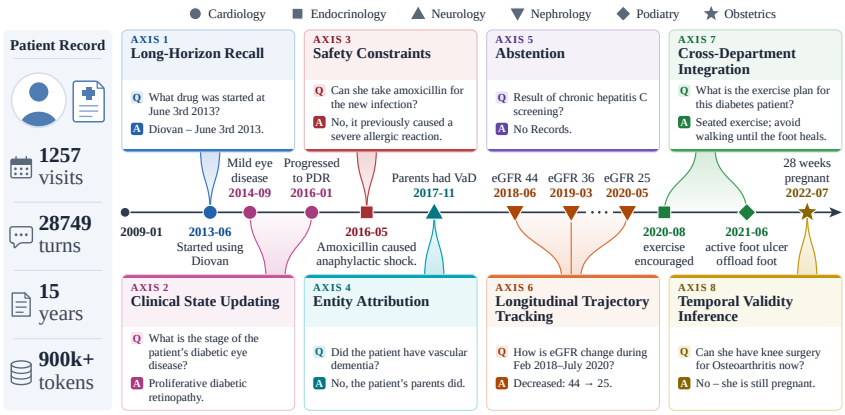


Figure 1: **Overview of the eight question axes in our benchmark** placed along a patient timeline representing a typical decade-long chronic-care patient record. We assign query time points of questions dynamically to ensure both large evidence-query distances and answer correctness.

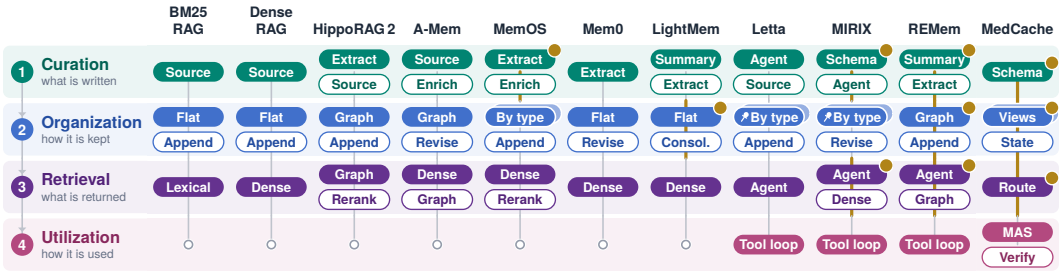
measured over years, recognize implicit status changes based on clinical observations, and disentangle information spread across multiple departments. For example, consider the development of a rehabilitation plan for a patient with diabetes. Such a plan depends on recommendations and restrictions documented by multiple departments over a span of several years, each contributing partial yet complementary information (Figure 1). If an agent's memory system fails to organize, retrieve, or utilize this information effectively, patient safety may be compromised, which in turn constrains the safe deployment of such systems in clinical practice. These limitations call for a reformulation of evaluation tasks that enables clinically grounded assessment.

Second, although concurrent medical benchmarks have begun to shift from single-visit to multi-visit assessment (Xu et al., 2026b; Zhang et al., 2026; Wang et al., 2026b), they remain insufficient to reveal how memory systems behave at the scale of real-world chronic care in practice. As shown in Figure 1, a patient with diabetes may accumulate more than a thousand visits across six departments over 15 years, yielding patient records of 900K+ tokens. In contrast, existing multi-visit medical benchmarks contain approximately 20 to 100 visits per patient, mostly drawn from limited number of years and departments Xu et al. (2026b); Zhang et al. (2026); Wang et al. (2026b). These gaps call for evaluation on substantially longer and realistic chronic care patient histories to assess whether memory systems can retain and utilize relevant information as records accumulate.

To close these gaps, we introduce **ME**mory systems in chronic **ME**dicine bench (MEME-Bench), a clinically grounded benchmark for evaluating and understanding agent memory systems in chronic care. Notably, MEME-Bench combines three key design principles not jointly offered by any existing benchmark, to our knowledge. First, it targets the demands of chronic care: its 1646 questions span eight capability axes, namely Long-Horizon Recall, Clinical State Updating, Safety Constraints, Entity Attribution, Abstention, Longitudinal Trajectory Tracking, Cross-Department Integration, and Temporal Validity Inference (Figure 1). Second, it matches the scale of real practice: its synthetic patient histories, grounded in real chronic-care hospital records, span 434–1718 visits across an average of 29.6 departments, far exceeding existing medical benchmarks. Third, it simulates clinical progression: memory systems ingest each history incrementally and answer questions at successive time points, using only the history observed so far. Leveraging MEME-Bench, we evaluate 11 memory systems and a long-context baseline across 7 backbone models to study four research questions: (1) How do existing memory systems perform in chronic care? (2) Which capabilities remain unresolved? (3) Where in the memory pipeline do failures arise? (4) Which design choices mitigate these failures?

Our main findings are as follows. (1) No memory system dominates across capabilities, and none outperforms the long-context baseline on every axis. (2) Longitudinal Trajectory Tracking and Cross-Department Integration remain poor for every system. To understand why, we trace the facts that each question requires from the patient record through the stored memory into the answering context, and switch the components of an agentic memory system on and off. Our results show that (3) retrieval is the most common source of failure: even systems that store every required fact often fail to bring it into the answering context, and these failures stem from challenges specific to chronic-care records; and (4) combining agentic retrieval with date-filtered search improves both

Table 1: Categorization of the evaluated memory systems, (| ●) marks a stage that uses event time.



of the hardest axes and lifts cross-department integration above every evaluated system, whereas prompting cannot substitute for these components.

In summary, our contributions are: ❶ **New Benchmark**: MEME-Bench evaluates agent memory on long, multi-department patient histories grounded in real chronic-care records, with questions on eight clinically motivated capabilities asked as the history accumulates. ❷ **New Diagnosis.** We characterize memory systems as a four-stage pipeline of curation, organization, retrieval and utilization, and trace the required facts through it to locate where each system fails. ❸ **New Findings.** We evaluate representative memory systems against a long-context baseline across backbones and, through controlled ablations, identify the design choices that matter most for chronic care.

2 Preliminaries: Four Stages of Memory Systems

Memory systems differ in many implementation choices, such as vector stores, knowledge graphs and agent tools, which obscures how they fundamentally differ in function. We therefore characterize each system as a pipeline of four stages (Table 1): *curation* decides what is written to memory; *organization* determines how it is stored and kept up to date; *retrieval* selects which stored information is returned for a given query; and *utilization* specifies how the retrieved evidence is used to generate an answer. We demonstrate the pipeline for a range of memory system categories, including BM25 and dense RAG baselines, HippoRAG 2 (Gutiérrez et al., 2025), A-Mem (Xu et al., 2026a), MemOS (Li et al., 2025), Mem0 (Chhikara et al., 2025), LightMem (Fang et al., 2026), Letta (Packer et al., 2023), MIRIX (Wang & Chen, 2025), REMem (Shu et al., 2026) and MedCache (Ting et al., 2026). 1 **Curation** either keeps verbatim units, such as raw chunks or messages Source (e.g., the RAG baselines), or transforms the input before writing it: an LLM extracts atomic facts or triples Extract (e.g., Mem0, HippoRAG 2), fills the fields of a predefined type system Schema (MIRIX, MedCache), condenses content into summaries or gists Summary (LightMem, REMem), or attaches generated metadata such as keywords and tags Enrich (A-Mem, MemOS). In some multi-store systems, the agent itself decides through tool calls what to write, when and where Agent (Letta, MIRIX). 2 **Organization** can be mainly categorized into single- and multi-store systems. A single store either keeps units unlinked, as in a vector database Flat (the RAG baselines, Mem0, LightMem), or connects them with explicit edges, as in a knowledge graph Graph (HippoRAG 2, A-Mem, REMem). Multi-store systems (stacked pills) split memory into one store per memory type, such as episodic and semantic memory By type (MemOS, Letta, MIRIX), or into overlapping specialty views of one shared store Views (MedCache). Some of these systems additionally maintain a core memory, such as persona and user profile information, that is always present in the context window and is updated directly by the agent; we denote this with a pin By type (Letta, MIRIX). As records accumulate, memory must also be maintained dynamically over time to handle superseded units: either append new units while leaving outdated ones untouched Append (e.g., the RAG baselines, Letta); update, merge or delete them online Revise (A-Mem, Mem0, MIRIX); consolidate them in offline batches Consol. (LightMem); or record an explicit validity or status for each unit State (MedCache). 3 **Retrieval** either ranks units in one shot by term overlap Lexical (BM25 RAG) or embedding similarity Dense (e.g., Dense RAG, Mem0), follows explicit edges at query time Graph (HippoRAG 2, A-Mem, REMem), lets the LLM issue and refine its own search calls over several rounds Agent (Letta, MIRIX, REMem), or sends the query to the relevant stores through a router Route (MedCache). A model may then rerank or filter the candidates Rerank (HippoRAG 2, MemOS). 4 **Utilization** either generates the answer in a single pass over the retrieved memory ○, interleaves reasoning with tool calls in a loop Tool loop (Letta, MIRIX, REMem), or splits

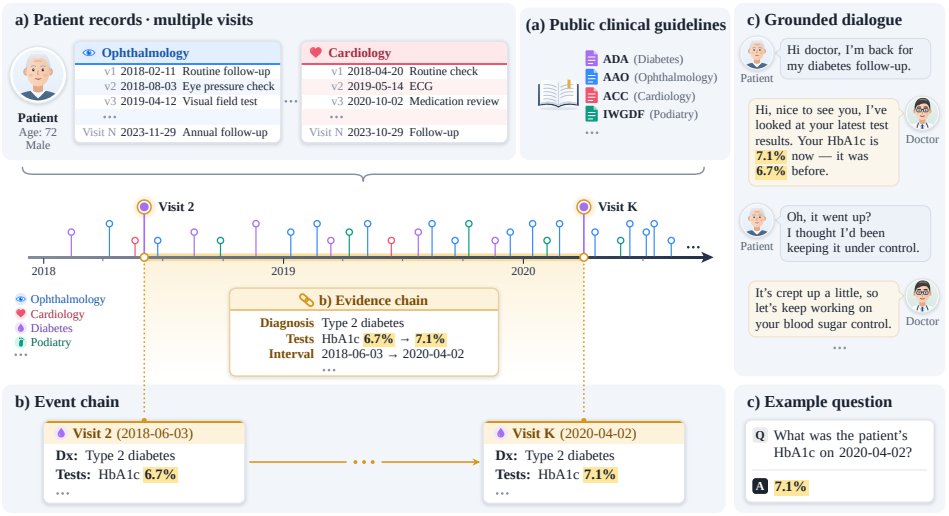


Figure 2: Construction pipeline of MEME-Bench.

reasoning across multiple agents MAS (MedCache); a system may also check the answer against memory before output Verify (MedCache).

3 BENCHMARK

MEME-Bench evaluates how memory systems retain and use patient information across long-term, multi-department clinical records. It includes histories for 23 synthetic patients, with a total of 15,049 visits and approximately 296,000 dialogue turns. Individual histories contain 434–1,718 visits, span a median of 14.7 years, and range from 150,000 to 954,000 tokens. MEME-Bench includes 1,646 questions across eight capability axes, asked at predefined visit checkpoints. Figure 2 illustrates the pipeline from real de-identified patient data to synthesized timelines, event and evidence chains, visit-level dialogues, and questions. Benchmark construction details of dataset statistics, de-identification, construction and manual verification are described in Appendix.

Data Sources and Synthetic Timeline Construction. We use real patient records to guide the longitudinal progression of the synthetic patient histories. The source records come from anonymized patients with chronic conditions at Anonymous Institution and include visit dates, departments, subjective, objective, assessment, and plan (SOAP) notes, diagnosis codes, and vital signs. Preprocessing includes de-identification, bilingual terminology normalization, visit-level segmentation, and extraction of numeric data from structured fields. Three human experts first abstract clinical measurements into trends and states and dates into relative intervals. An LLM then uses these structures to generate a sequence of synthetic clinical events across visits, with the numeric values synthesized under the guidance of publicly available clinical guidelines. This process produces synthetic visit-level timelines whose longitudinal structure is grounded in real clinical histories.

Event and Evidence Chain Construction. We convert each synthetic timeline into two complementary representations: a visit-level event chain for dialogue generation, and a fact-level evidence chain for constructing questions and reference answers that require single-fact retrieval or multi-visit reasoning. *Event chain:* An LLM decomposes each synthetic visit into atomic clinical facts organized under predefined clinical fields. Facts remain grouped by visit to preserve clinical context and temporal order, and each is assigned a unique identifier while retaining the dates, numerical values, medication names, and abbreviations. *Evidence chain:* Starting from the event chain, we group facts by department and use an LLM to link observations about the same clinical entity across visits, such as successive measurements or medication changes. Each linked record captures its values, effective intervals, status, and supporting visits, which can trace it back to the event chain. Three experts reviewed the evidence chains to verify that they faithfully represent the synthetic timeline.

Dialogue Generation. We convert each synthetic visit into a doctor–patient dialogue grounded in its event-chain facts. Two role-play agents receive different information. The doctor agent receives the current visit's facts, recent two visits in the same department, and a checklist of facts to com-

municate. The patient agent receives the presenting complaints, two recent same-department visits, and most recent visit to another department. Small talks are drawn from predefined non-clinical topics to introduce conversational noise without adding clinical information. Details such as prompt templates are in Appendix.

Task Axes. Clinicians managing chronic disease must review past investigations, reassess disease status, select medications, and coordinate care across specialties. We define eight capability axes to evaluate whether memory systems can well utilize the patient history needed to support routine clinical tasks: **(1) Long-Horizon Recall** tests recovery of specific historical facts, values, and findings at a given date, within an interval, or from a specified record. **(2) Clinical State Updating** tests whether memory systems can recover the state of a patient attribute supported by the most relevant record at the query time, while distinguishing it from earlier states. **(3) Safety Constraints** evaluates whether systems can retrieve and apply previously documented medication constraints such as allergies, when selecting among treatments, even if the question does not mention them explicitly. **(4) Entity Attribution** evaluates whether systems attribute clinical facts to the correct person, such as distinguishing the patient's conditions from family history, without losing information relevant to the patient's care. **(5) Abstention** tests whether systems acknowledge missing information rather than supply unsupported facts, especially when a question falsely presupposes that a test result, medication dose, or clinical state was documented. **(6) Longitudinal Trajectory Tracking** assesses whether systems can use dated clinical observations to compare time points, identify trends and extreme values, or locate the first documented abnormality. **(7) Cross-Department Integration.** Chronic care requires connecting findings and care plans across departments. For example, an ophthalmologist treating diabetic retinopathy needs to coordinate with the treatment plans documented in the patient's endocrinology records. This axis evaluates whether systems can combine such evidence across departments effectively, by determining whether requested investigations were completed and what they showed, whether prerequisites for a care plan were met, and whether active instructions conflict. **(8) Temporal Validity Inference.** Many clinical instructions remain valid only under certain conditions and lapse without an explicit update. For example, surgeries deferred during pregnancy may proceed after delivery. Acting on outdated instructions, or missing follow-ups they imply, is a well-documented safety concern (Callen et al., 2012). This axis evaluates whether systems can infer if an earlier instruction is still valid from relative durations and subsequent events.

Question Construction and Validation. We generate questions from the synthetic event and evidence chains using axis-specific templates, with reference answers drawn from selected facts. Each question is assigned a visit checkpoint that is distant from the nearest evidence required. Memory is built incrementally, and systems access only dialogues and visit metadata available through that checkpoint. For each question, we prompt GPT-5.6-Sol with and without the labeled gold evidence visits and record its answers. We then filter out questions (except Abstention axis) that LLM answers correctly without access to evidence, as these are answerable without memory, or incorrectly even with evidence, as these may be ambiguous or ill-posed. We additionally manually check answer–evidence consistency and temporal alignment.

4 RESULTS

Experimental Setup We evaluate Long Context baseline and 11 memory systems discussed in Sec. 2, Dense RAG, BM25 RAG, Hipporag 2, Letta, Remem, A-Mem, MemOS, LightMem, MIRIX, Mem0 and MedCache. Experiments are done on 7 backbones, including 1 close-sourced (GPT-5.1) and 6 open-sourced models (Qwen-3.5 0.8B, 4B, 9B, 27B, Qwen-3.6 27B, Gemma-4 31B). All numbers reported are average accuracy across three runs graded by GPT-5.6-Sol. Text exceeding backbone's max context window for long context baseline is truncated. Details including method adaptation, LLM-clinician consistency, latency and hyperparameter ablation are in Appendix.

4.1 RQ1 : HOW DO EXISTING MEMORY SYSTEMS PERFORM IN CHRONIC CARE?

MIRIX, Letta, REMem and MedCache lead overall on all three backbones consistently. As shown in Table 2, MIRIX and Letta lead the overall performance, followed by REMem and MedCache. Notably, these four are the only systems that outperform long-context baseline on Qwen-3.6 27B. The leaders are consistently strong on Gemma-4 31B and GPT-5.1 (Figure 3).

However, no method dominates every capability, and none outperforms long-context baseline on all eight axes. MIRIX is best on only two axes: Long-Horizon Recall and Safety Constraints.

Table 2: **Performance of 11 memory systems and long context across 8 axes** on Qwen 3.6 27B.

Method	Recall	State	Safety	Entity	Abst.	Traj.	X-dept.	Valid.	Mean
Long context	0.76	**0.79**	0.61	**0.88**	0.66	0.25	0.10	0.43	0.56
BM25 RAG	0.64	0.44	0.70	0.83	0.42	0.19	0.12	0.45	0.47
Dense RAG	0.61	0.61	0.70	0.85	0.54	0.22	0.13	0.37	0.50
HippoRAG 2	0.67	0.52	0.80	0.83	0.42	0.34	0.18	0.60	0.55
A-Mem	0.62	0.61	0.74	0.83	0.54	0.27	0.19	0.35	0.52
MemOS	0.67	0.48	0.83	0.81	0.45	0.26	0.17	0.50	0.52
Mem0	0.54	0.51	0.72	0.83	0.38	0.24	0.05	0.38	0.46
LightMem	0.70	0.51	0.69	0.82	0.41	0.26	0.11	0.31	0.48
Letta	0.74	**0.79**	0.89	0.73	0.82	0.47	0.45	0.75	0.71
MIRIX	**0.80**	0.76	**0.90**	0.68	0.89	0.52	0.47	0.70	**0.72**
REMem	0.73	0.55	0.83	0.69	0.92	**0.58**	0.19	**0.76**	0.66
MedCache	0.69	0.68	0.74	0.68	**0.99**	0.32	**0.51**	0.43	0.63

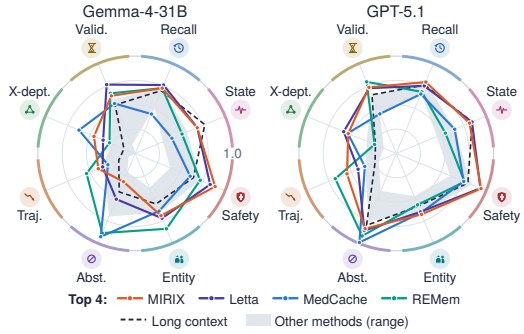


Figure 3: **Comparison of top 4 memory systems with long-context** on Gemma-4 31B and GPT-5.1 backbones. Grey: range of all other methods.

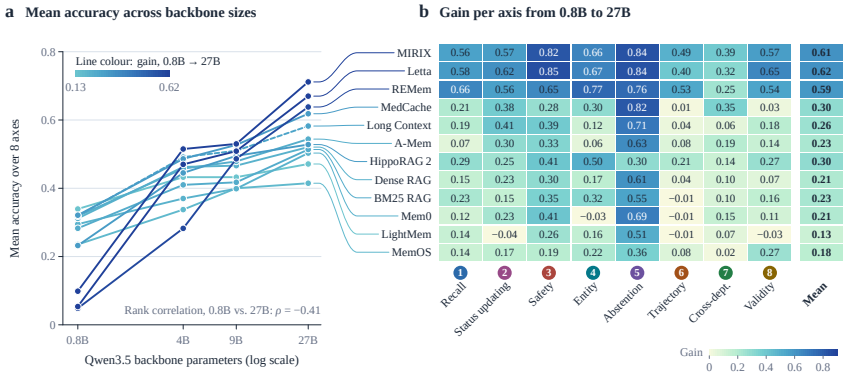


Figure 4: **Scaling the memory systems** across Qwen-3.5 0.8B, 4B, 9B and 27B backbones.

REMem, whose design includes date storage and date filtering during retrieval, leads Longitudinal Trajectory Tracking and Temporal Validity Inference. MedCache performs best on Abstention and Cross-Department Integration, while the long-context baseline leads on Entity Attribution and ties with Letta on Clinical State Updating. These top methods also show weak performance in Entity Attribution and inconsistent gains in Long-Horizon Recall, compared with long-context baseline.

Backbone scaling benefits memory systems unevenly across methods and axes. We next examine whether memory systems benefit from scaling up backbone models with enhanced capability. To test this, we evaluate all methods across Qwen-3.5 backbones of 0.8B, 4B, 9B and 27B parameters (Figure 4). Results show that the top methods (MIRIX, Letta, REMem) achieve the largest overall performance gains from backbone scaling. Despite weaker performance with a small backbone, these three methods benefit most from scaling up: their average accuracy gain is nearly three times that of the other methods. In contrast, six of the rest gain less than the long-context baseline. The top methods show decent gains across all axes, while the rest mainly benefit from the improvement on Abstention and Safety Constraint questions.

4.2 RQ2: What capabilities remain unresolved?

Longitudinal Trajectory Tracking and Cross-Department Integration are the two hardest axes across evaluations (Table 2). To understand what makes these two axes difficult, we characterize each along three dimensions (Figure 5a,b,d,e): the number of evidence visits required, the context length accumulated when it is asked, and its evidence depth, defined as the token distance to the supporting evidence. Each dot represents one question, positioned by its percentile rank on these dimensions and colored by local accuracy, computed by the mean accuracy of questions in its neighborhood in this three-dimensional space.

Cross-department accuracy falls as the evidence set grows. MIRIX is nearly insensitive to the context length and evidence depth, yet drops from 0.53 to 0.36 once four or more visits are required (Figure 5b, c), and similarly, other memory methods also decline with it.

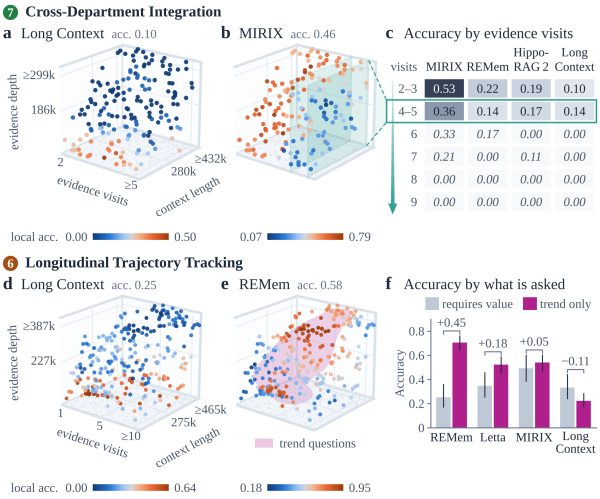


Figure 5: **Distribution of method performance** on the two hardest axes.

Table 3: **Oracle evidence supply analysis** on Qwen-3.6 27B. Detailed designs in Appendix.

Method	ΔOM (pp)	ΔOR (pp)
Letta	+11.3 [7.7, 14.5]	+18.5 [14.8, 22.1]
HippoRAG	+8.6 [4.9, 12.9]	+27.9 [23.6, 33.2]
REMem	+1.2 [−2.2, 5.6]	+25.3 [20.9, 30.7]
A-Mem	+16.1 [12.6, 19.4]	+27.9 [24.9, 31.8]
Dense	+16.3 [12.5, 20.5]	+29.4 [25.5, 34.6]
MemOS	+3.7 [0.3, 7.1]	+42.7 [38.6, 46.2]
BM25	−0.4 [−2.4, 1.5]	+37.1 [34.7, 39.5]
LightMem	+8.4 [4.5, 12.7]	+28.5 [24.8, 33.1]
MIRIX	−1.2 [−3.1, 0.7]	+9.1 [6.6, 11.6]
Mem0	+6.5 [4.4, 9.3]	+33.6 [30.9, 36.3]

Systems can identify trends yet miss extremes. Finding a historical minimum or maximum requires the recorded values and enough observations to compare, whereas a trend can be inferred from only a few. Accordingly, it is observed that every memory method answers trend-only questions more accurately than questions that require a value, except long-context baseline (Figure 5f). Long context fails with evidence depth instead: on both axes, its correct answers concentrate where the evidence lies close to the question (Figure 5a, d).

4.3 RQ3: Where in the memory pipeline do failures arise?

As discussed in Section. 2, given raw conversation records, memory systems typically first go through curation and organization to obtain and store memory items, and then apply retrieval based on questions to select relevant memory items, and finally reason with the augmented context. To investigate where and why memory systems fail, we trace annotated gold facts from the event and evidence chain into stored memory and the answering context. Figure. 6 shows the percentage of gold facts required for answering the question that are retained or retrieved at each stage of the memory system lifecycle, for six representative systems.

LLM-based memory construction without reference to raw notes often loses required facts. As shown in Figure 6, Dense RAG keeps every gold fact by storing the source chunks. HippoRAG 2 also keeps every gold fact, but mostly through its passages: the triples its LLM extracts hold only a small share of the required facts. While Mem0, LightMem, MIRIX and Letta all write memory items with LLMs, Mem0 and LightMem lose substantial amounts of gold evidence during construction and MIRIX a smaller share, while Letta preserves nearly all gold facts. A fact can be lost at either stage: it may never be written (curation), or be overwritten or merged later (organization). As shown in Table 1, Mem0 chooses to add new facts, delete obsolete facts or update the same fact with newer values over time, LightMem consolidates groups of similar facts into summaries, MIRIX routes each visit to managers that write typed records and revise them in place, and Letta appends written passages of newer visits and keeps them in an "archival store". In chronic care, an observation that has been superseded as the patient's current state may still be essential for identifying a historical extremum or reconstructing an earlier condition. In line with this, MIRIX holds far more of the required facts only in its time-stamped episodic memory than in its semantic memory of timeless facts. Mem0 and LightMem's update and consolidation lose important details. Letta suffers less by allowing access to raw data through its "conversation recall" tool.

Retrieval is the most common failure source. Despite retaining all required facts, Dense RAG is able to get only about 40% fully present in the answering context. Graph-based retrieval (HippoRAG 2) does not close this gap either. Mem0 and LightMem use the same similarity search and fully retrieve only about half of the facts they keep. Letta retrieves agentically, where it composes

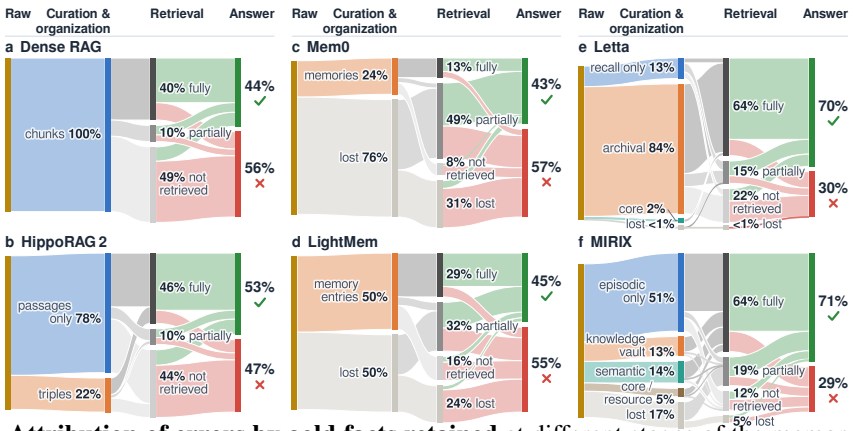


Figure 6: **Attribution of errors by gold facts retained** at different stages of the memory lifecycle for representative methods. Note that a fact lost doesn't mean the concept/visit is lost, so it still can be retrieved.

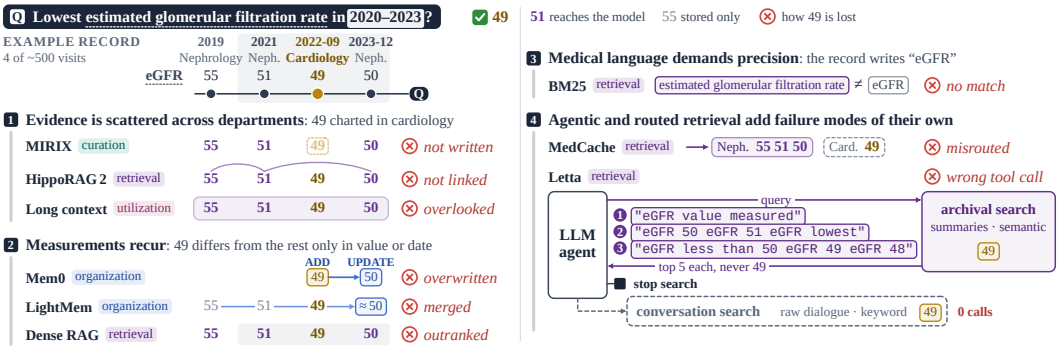


Figure 7: **Qualitative example** of same question but different failure modes across methods.

and reformulates its own search queries, chooses among its memory types (LLM-written archival passages vs. raw message recall), and decides when it has enough to answer, bringing more of the required facts fully into context than any single-pass method. The other top methods also let a model decide how to retrieve: MIRIX and REMem search agentically and can restrict a search to a time range (MIRIX in its episodic store, REMem in its episodic graph), while MedCache routes each question to specialty views with a trained router.

These losses stem from unique challenges in chronic-care records; agentic and routed retrieval add failure modes of their own. We qualitatively attribute them to three such challenges, which Figure 7 traces through a single question that, in this illustration, every method answers wrongly. (1) *Evidence is scattered across departments.* A patient with several chronic conditions is seen by several departments, and a value needed later is often charted in passing by a department other than the one that tracks it: the lowest eGFR in the window was charted at a cardiology visit, not at the nephrology visits. MIRIX's managers write each visit up around its main topic and never store that reading (curation); HippoRAG 2's graph links the nephrology visits to one another but not to the cardiology visit, so its walk never reaches it (retrieval). (2) *Measurements recur*, so the required reading differs from many near-identical ones only in value or date. This is why updating memory erases what extremum questions ask about, as when Mem0 overwrites 49 with a later 50 and LightMem merges the readings into one entry, and why similarity search, blind to value and date, spends its top five on similar readings from other years, so that only some readings in the window reach Dense RAG's question-answering LLM. (3) *Medical language demands precision*: BM25 cannot match the question's "estimated glomerular filtration rate" to the record's "eGFR", so no eGFR visit reaches its reader at all. Agentic and routed retrieval add failure modes of their own, since each decision about where, how or how long to search can go wrong: MedCache routes the question to the nephrology view, missing the reading charted in cardiology, and Letta keeps the minimum in both its archival memory and its raw dialogue, yet searches only the archival, writes a numeric condition into a similarity query, and stops without a keyword search.

Finally, we compare each method's original performance with two interventions as shown in Table. 3: supplying gold evidence in memory storages (OM) and supplying it directly in the retrieval returns (OR). We observe that OR improves accuracy in every method, while OM improvement intervals for ReMem, BM25, and MIRIX include zero. OR also yields higher point estimates than OM throughout the table. These results further verify the retrieval gaps observed in Figure. 6.

4.4 RQ4: Which design choices mitigate these failures?

As described in RQ3, methods differ in several design choices at once, so their scores cannot tell which choice matters. Therefore, to isolate each component, we start from Letta, the simplest of the four, and switch its components on and off (Table 4). We remove its core-memory block (does keeping a working memory of the patient help?), its access to the raw dialogue (can the verbatim record recover what curation drops?) and its agentic retrieval, replaced by single-shot dense retrieval over the same stores (does it matter that the model plans and refines its own searches?). We add the temporal designs of REMem and MedCache: visit dates on every memory item (does knowing when each fact was observed help?) and a date filter on search (can retrieval act on those dates?). Finally, we try prompts in place of components: date and evidence instructions at answering time (can instructions do the work of the components?), and a write prompt that dates every entry as memory is built (can date at curation time replace filtering at search time?).

Agentic retrieval matters most, and access to the raw dialogue adds a smaller gain. Replacing agentic retrieval costs far more than any other change and brings Letta down to the level of Dense RAG, so its lead in RQ1 comes mostly from how it retrieves rather than from what it stores. Access to the raw dialogue helps in every matched pair, consistent with observations in RQ3, where some required facts survive only there. The core-memory block makes little difference.

Dates help only when search can filter on them. Visit dates added during memory construction alone change little, but adding a date filter improves every matched pair and gives the best run. The filter lets retrieval select recurring readings by date, which similarity search cannot. Its largest gain is on Cross-Department Integration, which no method solved in RQ2: every filtered run surpasses the best method in Table 2.

Prompting does not fully substitute for these components. Date and evidence instructions alone do not offer large accuracy gains. When combined, they obtain offer slight improvements over Letta baseline, while prompted runs overflow the context far more often than the filtered run and take longer per question. A write prompt that requires adding dates and attributes every entry does not help.

Table 4: **Which design choices matter:** runs of Letta and its variants on Qwen-3.6 27B.

Loop	Dates	Filter	Raw	Core	Prompt: Date	Prompt: Evid.	Prompt: Write	Overall	Δ vs. Letta	Traj.	X-dept.
○	○	○	●	●	○			.505	**−.201**	.24	.15
○	●	○	●	●	○			.518	**−.188**	.27	.11
●	○	○	●	●	○			.706	*Letta*	.47	.45
●	○	○	●	○	○			.709	+.003	.46	.45
●	○	○	○	●	○			.693	−.013	.44	.37
●	○	○	○	○	○			.666	**−.040**	.42	.36
●	●	○	●	●	○			.716	+.010	.54	.37
●	●	○	●	○	○			.697	−.009	.49	.45
●	●	○	○	●	○			.682	−.024	.42	.38
●	●	○	○	○	○			.669	**−.037**	.40	.32
●	●	●	●	●	○			.736	**+.030**	.53	.63
●	●	●	●	○	○			.729	**+.023**	.53	.60
●	●	●	○	●	○			.701	−.005	.48	.56
●	●	●	○	○	○			.707	+.001	.53	.58
●	○	○	●	●	●			.720	+.014	.56	.45
●	●	○	●	●	●			.696	−.010	.40	.38
●	●	○	●	●	●	●		.727	**+.021**	.51	.45
●	○	○	●	●			●	.687	−.019	.42	.40
●	○	○	●	●	●	●	●	.690	−.016	.48	.38

5 Conclusion

We introduce MEME-Bench to evaluate and understand agent memory systems in chronic care. We find that: no evaluated system outperforms long context baseline across all eight evaluated abilities, and all struggle to track trajectories and to integrate evidence across departments; retrieval is the most common source of failure, as stored facts often never reach the answering model; agentic retrieval with date-filtered search is the most helpful design choice, while prompting alone can not replace these components. We hope MEME-Bench guides the design of memory systems in future chronic care research.

AI USE STATEMENT

We used AI to generate synthetic data sets, implement methods, clean and reformat dataset, support data analysis, interpret results, writing and figure drawing, identify relevant literature, and propose the benchmark name.

We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS

This is a purely synthetic medical dataset. We used guidance from real datasets which underwent rigorous de-identification under institution ethics approval. Details can be found in Appendix.

REPRODUCIBILITY STATEMENT

We specify details of the benchmark and results in paper and Appendix. We publicly host a project page on huggingface and the benchmark, code and results will be released.

REFERENCES

Introduction and methodology: Standards of care in diabetes—2026, 2026.

Robert H Brophy and Yale A Fillingham. Aaos clinical practice guideline summary: management of osteoarthritis of the knee (nonarthroplasty). *Journal of the American Academy of Orthopaedic Surgeons*, 30(9):e721–e729, 2022.

Joanne L Callen, Johanna I Westbrook, Andrew Georgiou, and Julie Li. Failure to follow-up test results for ambulatory patients: a systematic review. *Journal of general internal medicine*, 27 (10):1334–1348, 2012.

Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman. When helpfulness backfires: Llms and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8(1):605, 2025.

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

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Agency for Healthcare Research and Quality. Chronic condition indicator refined (ccir) for icd-10-cm, 2022.

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

Alan E Guttmacher, Francis S Collins, and Richard H Carmona. The family history—more important than ever. *New England Journal of Medicine*, 351(22):2333–2336, 2004.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145 (18):e895–e1032, 2022.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Eva S Klappe, Jarom Heijmans, Kaz Groen, Judith Ter Schure, Ronald Cornet, and Nicolette F de Keizer. Correctly structured problem lists lead to better and faster clinical decision-making in electronic health records compared to non-curated problem lists: A single-blinded crossover randomized controlled trial. *International journal of medical informatics*, 180:105264, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Jiajia Li, Zhouyu Guan, Jing Wang, Carol Y Cheung, Yingfeng Zheng, Lee-Ling Lim, Cynthia Ciwei Lim, Paisan Ruamviboonsuk, Rajiv Raman, Leonor Corsino, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nature medicine*, 30(10):2886–2896, 2024.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, et al. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*, 2025.

Jennifer I Lim, Stephen J Kim, Steven T Bailey, Jaclyn L Kovach, G Atma Vemulakonda, Guishuang Ying, and Christina J Flaxel. Diabetic retinopathy preferred practice pattern®. *Ophthalmology*, 132(4):P75–P162, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

American College of Obstetricians, Gynecologists, et al. Acog committee opinion no. 775: nonobstetric surgery during pregnancy. *Obstet Gynecol*, 133(4):e285–e286, 2019.

World Health Organization et al. World health organization noncommunicable diseases. *World Health Organization (2018, June 1) Retrieved from https://www. who. int/news-room/fact-sheets/detail/noncommunicable-diseases*, 2018.

J Marc Overhage and David McCallie Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Annals of internal medicine*, 172(3):169–174, 2020.

Ann S. O'Malley and James D. Reschovsky. Referral and consultation communication between primary care and specialist physicians: Finding common ground. *Archives of Internal Medicine*, 171(1):56–65, 01 2011. ISSN 0003-9926. doi: 10.1001/archinternmed.2010.480. URL `https://doi.org/10.1001/archinternmed.2010.480`.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Hoangmai H Pham, Deborah Schrag, Ann S O'Malley, Beny Wu, and Peter B Bach. Care patterns in medicare and their implications for pay for performance. *New England Journal of Medicine*, 356(11):1130–1139, 2007.

Justin Porter, Cynthia Boyd, M Reza Skandari, and Neda Laiteerapong. Revisiting the time needed to provide adult primary care: Porter et al. *Journal of general internal medicine*, 38(1):147–155, 2023.

Justin J Sanders, Sarah Temin, Arun Ghoshal, Erin R Alesi, Zipporah Vunoro Ali, Cynthia Chauhan, James F Cleary, Andrew S Epstein, Janice I Firn, Joshua A Jones, et al. Palliative care for patients with cancer: Asco guideline update. *Journal of Clinical Oncology*, 42(19):2336–2357, 2024.

Jaspreet S Sandhu, Brooke R Bixler, Philipp Dahm, Ramy Goueli, Erin Kirkby, John T Stoffel, and Timothy J Wilt. Management of lower urinary tract symptoms attributed to benign prostatic hyperplasia (bph): Aua guideline amendment 2023. *The Journal of urology*, 211(1):11–19, 2024.

Éric Senneville, Zaina Albalawi, Suzanne A Van Asten, Zulfiqarali G Abbas, Geneve Allison, Javier Aragón-Sánchez, John M Embil, Lawrence A Lavery, Majdi Alhasan, Orhan Oz, et al. Iwgdf/idsa guidelines on the diagnosis and treatment of diabetes-related foot infections (iwgdf/idsa 2023). *Diabetes/metabolism research and reviews*, 40(3):e3687, 2024.

Feichen Shen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 national nlp clinical challenges (n2c2)/open health natural language processing (ohnlp) competition. *JMIR medical informatics*, 9(1):e24008, 2021.

Yiheng Shu, Padmaja Jonnalagedda, Xiang Gao, Bernal Jimenez Gutierrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. Remem: Reasoning with episodic memory in language agent. In *International Conference on Learning Representations*, volume 2026, pp. 55958–55987, 2026.

Søren T Skou, Frances S Mair, Martin Fortin, Bruce Guthrie, Bruno P Nunes, J Jaime Miranda, Cynthia M Boyd, Sanghamitra Pati, Sally Mtenga, and Susan M Smith. Multimorbidity. *Nature reviews Disease primers*, 8(1):48, 2022.

Peter C Smith, Rodrigo Araya-Guerra, Caroline Bublitz, Bennett Parnes, L Miriam Dickinson, Rebecca Van Vorst, John M Westfall, and Wilson D Pace. Missing clinical information during primary care visits. *Jama*, 293(5):565–571, 2005.

Paul E Stevens, Sofia B Ahmed, Juan Jesus Carrero, Bethany Foster, Anna Francis, Rasheeda K Hall, Will G Herrington, Guy Hill, Lesley A Inker, Rümeyza Kazancıoğlu, et al. Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international*, 105(4):S117–S314, 2024.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. In *International Conference on Learning Representations*, volume 2026, pp. 133343–133429, 2026.

Hei Ting, Chenwei Wu, Xueshen Liu, Boyuan Zheng, Liyue Shen, Jiasi Chen, Z Morley Mao, et al. Medcache: Efficient and temporally valid memory for longitudinal clinical agents. *arXiv preprint arXiv:2608.29528*, 2026.

National Clinical Guideline Centre UK. Drug allergy: diagnosis and management of drug allergy in adults, children and young people. 2014.

Huimin Wang, Zhengyi Zhao, and Yutian Zhao. Clintracebench: Source-verifiable longitudinal clinical reasoning over ehr-derived dialogues. *arXiv preprint arXiv:2609.01111*, 2026a.

Yihao Wang, Haoran Xu, Renjie Gu, Yixuan Ye, Xinyi Chen, Xinyu Mu, Yuan Gao, Chunxiao Guo, Peng Wei, Jinjie Gu, et al. Medmemorybench: Benchmarking agent memory in personalized healthcare. *arXiv preprint arXiv:2605.11814*, 2026b.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Lawrence L Weed. Medical records that guide and teach. In *Clinical Problem Lists in the Electronic Health Record*, pp. 39–64. Apple Academic Press, 2019.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026a.

Zihan Xu, Yanzhen Chen, Xiaocheng Zhang, Zhiting Fan, Weiqi Zhai, Hongxia Xu, and Zuozhu Liu. Longmedbench: Benchmarking medical agents for long-horizon clinical decision-making. *arXiv preprint arXiv:2607.09322*, 2026b.

Clyde W Yancy, Mariell Jessup, Biykem Bozkurt, Javed Butler, Donald E Casey Jr, Monica M Colvin, Mark H Drazner, Gerasimos S Filippatos, Gregg C Fonarow, Michael M Givertz, et al. 2017 acc/aha/hfsa focused update of the 2013 accf/aha guideline for the management of heart failure: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart failure society of america. *Circulation*, 136(6):e137–e161, 2017.

Zeyu Zhang, Ziqing Wang, and Kaize Ding. Medlocomo: A long-context multi-session medical dialogue benchmark for large language models. *arXiv preprint arXiv:2607.22566*, 2026.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, et al. Ama-bench: Evaluating long-horizon memory for agentic applications. *arXiv preprint arXiv:2602.22769*, 2026.