Beyond Retrieval: Bi-Temporal State Arbitration for Longitudinal Healthcare Agents

Jianing Zhao
Tianjin University
zhaojianing@tju.edu.cn

Xiaoquan Zhi
Tianjin University
zhixiaoquan@tju.edu.cn

Xinqiang Yu
Tianjin University
yu2651701064@tju.edu.cn

Abstract

Longitudinal healthcare agents require persistent state tracking under temporal uncertainty. In domains like chronic disease management, patient states—medications, symptoms, and vital signs—evolve continuously over months. Existing memory architectures for Large Language Models (LLMs) are inherently *retrieval-centric*: they treat memory as a static repository of past interactions, failing to resolve conflicting or superseded information when queried for the current patient state. We propose a shift to *state-centric* memory. Our framework introduces (1) a bi-temporal state representation that decouples event time from ingestion time and tracks temporal validity windows, (2) an incremental state arbitration mechanism using four operators—SUPPORT, REFINE, SUPERSEDE, and BRANCH-CONFLICT—to handle evolving medical facts without destructive overwriting, and (3) a confidence-thresholded evidence escalation layer for robust, efficient memory access. Evaluated on a longitudinal diabetes management suite as a representative biomedical state tracking task, our method achieves a Unique-F1 of 0.85 and Conflict-F1 of 0.98, substantially improves upon long-context LLMs (0.38 / 0.89) and standard vector memory (0.30 / 0.60), demonstrating that agentic AI in longitudinal biomedical settings requires continuous, evidence-grounded arbitration rather than simple retrieval.

1 Introduction

The application of Large Language Models (LLMs) in healthcare has rapidly advanced from static question answering (Singhal et al., 2025) to the development of autonomous healthcare agents capable of multi-step clinical reasoning (Gao et al., 2024; Ge et al., 2026). However, a critical architectural gap remains: longitudinal healthcare agents require *persistent state tracking under temporal uncertainty*. In real-world chronic disease management—diabetes being a canonical example—a patient's clinical state is never static. Medications are initiated, adjusted, and discontinued; symptoms emerge and resolve; and vital sign trajectories shift across days and weeks (Li et al., 2025).

When an LLM agent interacts with a patient over months, the same attribute slot (e.g., "metformin dosage") will be mentioned multiple times. Subsequent mentions may be supplementary, corrective, substitutive, or entirely contradictory to previous records. Consider the following motivating example from a longitudinal diabetes management scenario:

> **Turn 3 (Jan 10):** *Patient:* "I take Metformin 500mg every morning."
> **Turn 11 (Feb 14):** *Patient:* "My doctor increased my dosage to 1000mg."
> **Turn 22 (Mar 05):** *Patient:* "I switched to Insulin 10U three days ago because the pills upset my stomach."
> **Turn 25 (Mar 08):** *Patient:* "Should I take my 1000mg pill before breakfast?"

In this scenario, a naive retrieval system might fetch Turn 11 and incorrectly advise the patient to take the 1000mg pill, ignoring the subsequent switch to Insulin (Turn 22) or failing to recognize that Turn 25 is a conflicting statement that requires clarification. The agent must not only *remember* these mentions but also *arbitrate* among them to maintain a coherent, current picture of the patient's health state.

Existing memory architectures for LLM agents are fundamentally ill-suited for this task. Standard vector-based Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) retrieves the most *similar* past fragments, not the most *currently valid* state. Graph-based systems like Graphiti/Zep (Rasmussen et al., 2025) construct temporal knowledge graphs but still rely on the LLM to resolve conflicts among retrieved nodes at inference time. Long-context windows suffer

from the "lost in the middle" phenomenon and scale poorly with session length. Crucially, all of these approaches are *retrieval-centric*: their goal is to find relevant past information, not to maintain a living, arbitrated model of the present.

We propose a paradigm shift from retrieval-centric to **state-centric** memory for longitudinal healthcare agents. The core insight is that the agent's primary memory task is not "find what was said" but rather "maintain what is currently true." This reframing draws on two classical traditions: Truth Maintenance Systems (TMS) from AI (Doyle, 1979), which formalize belief revision under new evidence, and bi-temporal databases from the database community (Jensen et al., 1994), which distinguish between when a fact was recorded and when it was actually true.

Our contributions are threefold:

- **Bi-Temporal Patient State Representation:** A structured state unit that decouples event time from ingestion time and explicitly tracks temporal validity windows, enabling accurate retroactive reporting and historical state reconstruction.
- **Incremental State Arbitration:** A four-operator mechanism (SUPPORT, REFINE, SUPERSEDE, BRANCH-CONFLICT) that updates the state graph incrementally as new evidence arrives, preserving conflict branches rather than destructively overwriting.
- **Confidence-Calibrated Evidence Escalation:** A tiered access policy that routes queries across memory layers based on calibrated confidence thresholds, minimizing latency while maintaining answer quality.

2 Related Work

LLM Agents in Healthcare. Recent work has explored multi-agent frameworks and autonomous LLMs for clinical decision-making (Ge et al., 2026; Liu et al., 2025). Systems like Med-PaLM (Singhal et al., 2025) and CARE-AD (Li et al., 2025) demonstrate strong reasoning capabilities in medical contexts. However, most existing clinical agents focus on single-turn interactions or short diagnostic sessions. They lack a persistent, evolving state maintenance mechanism required for longitudinal tracking over months, typically relying on the LLM's context window, which is both expensive and unreliable for long-horizon state tracking.

Memory-Augmented LLM Agents. To extend LLMs with long-term memory, MemGPT (Packer et al., 2023) and Mem0 (Chhikara et al., 2025) manage hierarchical memory tiers with explicit read/write operations. Memory OS (Kang et al., 2025) introduces an operating-system metaphor for agent memory management. Graph-based approaches, including Graphiti and Zep (Rasmussen et al., 2025; Yang et al., 2026), construct temporally-aware knowledge graphs. Despite their sophistication, these systems remain retrieval-centric: when confronted with conflicting evidence across time, they either return all retrieved nodes to the LLM for in-context resolution or apply naive overwrite updates. Neither approach maintains a persistent, arbitrated state that can be directly queried for the current truth.

Belief Revision and Bi-Temporal Databases. The problem of maintaining consistent beliefs under new evidence is foundational in AI, formalized by Doyle's Truth Maintenance System (Doyle, 1979) and the belief revision framework of Gärdenfors (Gärdenfors, 1988). Independently, the database community developed bi-temporal data models (Snodgrass and Ahn, 1985; Jensen et al., 1994) to track both transaction time (when data was recorded) and valid time (when a fact held in reality). Our work operationalizes these classical concepts for natural-language-driven healthcare agents, adapting them to handle the soft confidence scores and linguistic ambiguity inherent in LLM-mediated interactions. We are not introducing bi-temporality or belief revision as novel concepts; rather, we operationalize these classic concepts for incremental, natural-language-driven healthcare agents, adapting them to handle the soft confidence and ambiguity inherent in LLM interactions. This represents a fundamental architectural departure from current memory-augmented generation systems, which typically treat all retrieved text as equally valid assertions of fact. By explicitly modeling the difference between a patient's historical report and their current physiological reality, our framework bridges the gap between conversational AI and formal medical informatics.

3 Bi-Temporal Patient State Memory

In a state-centric paradigm, the fundamental unit of memory is not a conversational turn or a raw text chunk, but a structured belief about the patient's health status. We define this unit as the `CanonicalMemory` object, which serves as the primary node in the patient state graph.

3.1 The Bi-Temporal Model

Standard temporal knowledge graphs associate a single timestamp with an edge, conflating the time a fact was recorded with the time it was true. This conflation is particularly problematic in medical dialogues, where patients routinely report past events retroactively (e.g., "I stopped taking Metformin three days ago, but forgot to mention it"). Our state unit explicitly tracks four distinct temporal fields:

- t_{event}: The calendar time the clinical event actually occurred in the real world.
- t_{ingest}: The logical time the system learned of the event, represented by the dialogue turn number τ.
- $t_{\text{valid_start}}$: The beginning of the temporal window during which the medical fact is considered true.
- $t_{\text{valid_end}}$: The end of the temporal window. A value of ∞ (or `None`) indicates the state is currently ongoing.

This bi-temporal separation enables the agent to correctly handle retroactive reporting. As illustrated in Figure 1, when a patient reports at turn τ_{ingest} that an event occurred at $t_{\text{event}} < t_{\text{ingest}}$ (e.g., switching to Insulin "three days ago"), the system inserts the state unit at the correct position in the valid-time timeline without corrupting the ingestion-time record of when the information was received.

Formally, a state unit is a tuple:

$$\mathcal{S} = \langle \text{slot}, \text{candidates}, t_{\text{event}}, t_{\text{ingest}}, t_{\text{valid_start}}, t_{\text{valid_end}}, \text{status} \rangle \quad (1)$$

where `slot` identifies the attribute (e.g., *medication.metformin.dosage*), `candidates` is a ranked list of candidate values, and `status` records whether the state is active, superseded, conflicting, or resolved.

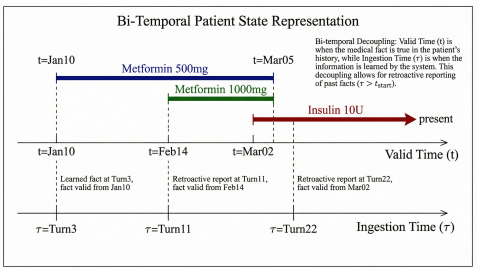

Figure 1: Bi-Temporal Patient State Representation. The valid-time axis (t) tracks when a state is medically true, while the ingestion-time axis (τ) tracks when the agent learned of it. Retroactive reporting (e.g., Turn 22) requires updating the valid-time timeline retroactively.

3.2 Confidence-Weighted Candidate Values

In real-world medical dialogues, the same attribute slot may receive conflicting reports. Rather than prematurely committing to a single value, our state unit maintains a list of `CandidateValue` objects. Each candidate v_i is associated with a composite confidence score:

$$c_i = \alpha \cdot w_{\text{auth}}(v_i) + \beta \cdot w_{\text{recency}}(v_i) + \gamma \cdot f(n_{\text{evid}}(v_i)) \quad (2)$$

where $w_{\text{auth}} \in [0, 1]$ encodes source authority (clinician directive: 1.0; patient self-report: 0.5; inferred: 0.3), $w_{\text{recency}} = \tau / \tau_{\text{max}}$ normalizes by the global maximum turn, n_{evid} counts independent corroborating mentions, and $f(\cdot) = \log(1 + n_{\text{evid}})$ is a diminishing-returns function. The hyperparameters α, β, γ are set to $0.5, 0.3, 0.2$ respectively, calibrated on a held-out validation set. The candidate with the highest c_i is designated the *representative value* of the state unit.

This representation upgrades conflict resolution from a binary "pick the winner" operation to a continuous "rank by evidence" process, preserving uncertainty and enabling future arbitration as new evidence arrives.

4 Incremental State Arbitration

When the agent receives new evidence E_{new} extracted from a dialogue turn, it does not simply overwrite the existing state. Instead, it invokes the **State Arbitration** procedure, which evaluates the relationship between E_{new} and the currently active state unit $\mathcal{S}_{\text{old}}$ for the same slot, and applies one of four operators.

4.1 The Four Arbitration Operators

As illustrated in Figure 2, the arbitration mechanism evaluates the relationship between E_{new} and the currently active state unit $\mathcal{S}_{\text{old}}$, applying one of four operators:

Operator	Trigger	State Update
SUPPORT	Same value	Increase confidence; add provenance
REFINE	Adds detail	Merge qualifiers; keep history
SUPERSEDE	Explicit replacement	Close old valid window; activate new state
BRANCH-CONFLICT	Contradiction without replacement	Keep competing candidates with confidences

Figure 2: Four State Arbitration Operators. New evidence triggers one of four operators, updating the confidence scores, validity windows, or candidate lists of the patient state unit.

SUPPORT. Triggered when the value of E_{new} matches the representative value of $\mathcal{S}_{\text{old}}$ within a semantic equivalence threshold. The new evidence corroborates the existing belief. The system increments n_{evid} of the matching candidate, appends the provenance of E_{new}, and recomputes c_i via Equation 2. The validity window is unchanged.

REFINE. Triggered when E_{new} adds specificity to $\mathcal{S}_{\text{old}}$ without contradiction (e.g., updating "Metformin" to "Metformin 500mg twice daily"). The system merges the new qualifiers into the existing active state while preserving historical provenance. The representative value is updated in place.

SUPERSEDE. Triggered when E_{new} explicitly replaces $\mathcal{S}_{\text{old}}$, identified via lexical action patterns (e.g., "stopped taking", "switched to", "no longer"). The system sets $t_{\text{valid_end}}$ of $\mathcal{S}_{\text{old}}$ to $t_{\text{event}}(E_{\text{new}})$ and transitions its status to SUPERSEDED. A new state unit $\mathcal{S}_{\text{new}}$ is created with $t_{\text{valid_start}} = t_{\text{event}}(E_{\text{new}})$ and status ACTIVE. This creates a continuous, non-overlapping temporal chain of states for the slot.

BRANCH-CONFLICT. Triggered when E_{new} contradicts $\mathcal{S}_{\text{old}}$ but lacks explicit supersede intent. The system adds a new `CandidateValue` to $\mathcal{S}_{\text{old}}$'s candidate list and recomputes all confidence scores. The status of $\mathcal{S}_{\text{old}}$ transitions to CONFLICTING. The conflict is recorded in a `ConflictRecord` with full provenance, enabling downstream explanation and future resolution.

Algorithm 1 State Arbitration

Input: New evidence E_{new}, state graph $\mathcal{G}$
Output: Updated state graph $\mathcal{G}'$
$\mathcal{S}_{\text{old}} \leftarrow \text{QueryActiveState}(\mathcal{G}, E_{\text{new}}.\text{slot})$
if $\mathcal{S}_{\text{old}} = \emptyset$ **then**
  $\mathcal{G}' \leftarrow \mathcal{G} \cup \text{CreateState}(E_{\text{new}})$
else if $\text{IsSupersede}(E_{\text{new}})$ **then**
  $\mathcal{S}_{\text{old}}.t_{\text{valid_end}} \leftarrow E_{\text{new}}.t_{\text{event}}$
  $\mathcal{S}_{\text{old}}.\text{status} \leftarrow$ SUPERSEDED
  $\mathcal{G}' \leftarrow \mathcal{G} \cup \text{CreateState}(E_{\text{new}})$
else if $\text{IsMatch}(E_{\text{new}}, \mathcal{S}_{\text{old}})$ **then**
  $\text{ApplySupport}(\mathcal{S}_{\text{old}}, E_{\text{new}})$; $\mathcal{G}' \leftarrow \mathcal{G}$
else if $\text{IsRefinement}(E_{\text{new}}, \mathcal{S}_{\text{old}})$ **then**
  $\text{ApplyRefine}(\mathcal{S}_{\text{old}}, E_{\text{new}})$; $\mathcal{G}' \leftarrow \mathcal{G}$
else
  $\text{ApplyBranchConflict}(\mathcal{S}_{\text{old}}, E_{\text{new}})$
  $\text{ApplyConstraints}(\mathcal{S}_{\text{old}})$; $\mathcal{G}' \leftarrow \mathcal{G}$
end if

4.2 Domain-Specific Formal Constraints

To ensure clinical coherence without costly LLM calls on every update, the arbitration mechanism applies lightweight rule-based constraints after the BRANCH-CONFLICT operator. These constraints adjust candidate confidence scores to penalize clinically implausible states. Examples include: (1) a *MedicationStartStopConstraint* that penalizes candidates implying a medication was both started and stopped on the same day; (2) a *DiagnosisTestOrderConstraint* that penalizes diagnostic conclusions that precede the corresponding test results; and (3) a *DosageMonotonicityConstraint* that flags non-monotonic dosage changes without an explicit clinical rationale. These constraints encode domain knowledge as confidence penalties, allowing the system to prefer clinically coherent interpretations without requiring a full reasoning pass.

5 Confidence-Thresholded Evidence Escalation

The state arbitration mechanism ensures the accuracy of the memory graph. However, efficiently accessing this information during real-time dialogue requires a structured query policy. We propose a **confidence-thresholded evidence escalation** layer that routes queries across memory tiers based on the confidence of available evidence, balancing computational cost with answer robustness.

The escalation policy follows a four-tier hierarchy. At each tier, the system evaluates whether the available evidence is sufficient to answer the query. If so, it returns immediately; otherwise, it escalates to the next tier.

Tier 1 (State Check): The system queries the active `CanonicalMemory` states for the relevant slot. If the representative candidate confidence $c^* > \theta_{\text{low}}$, the answer is returned directly from the state unit. This tier is $O(1)$ and represents the "current truth" as maintained by the arbitration mechanism.

Tier 2 (Structured Query): If the query involves numerical aggregation (e.g., "average blood glucose over 7 days") or the state confidence is below θ_{low} for a quantitative slot, the system issues a structured query to the SQLite vitals backend. This tier provides exact computations that LLMs cannot reliably perform.

Tier 3 (Episodic Fallback): If neither Tier 1 nor Tier 2 is sufficient (e.g., the query requires conversational context or the state is CONFLICTING), the system retrieves the top-K episodic fragments via dense vector retrieval. This tier provides the raw conversational evidence underlying the state.

Tier 4 (Graph Traversal): If the top-1 episodic similarity falls below θ_{vec}, the system escalates to graph traversal, exploring cross-slot causal links for multi-hop queries (e.g., "Why did my doctor change my medication?"). This is the most expensive tier and is invoked only when lower tiers are insufficient.

The thresholds $\theta_{\text{low}} = 0.65$ and $\theta_{\text{vec}} = 0.72$ were determined empirically by grid search on a held-out validation split (10 dialogues from the "medium" difficulty set), optimizing for the harmonic mean of QA accuracy and mean access latency. The high proportion of queries (71.4%) resolved at Tier 1 is due to the query mix in our evaluation suite, where the majority of patient questions target their current, active state rather than historical context.

6 Experiments

6.1 Experimental Setup

We evaluate our framework on a longitudinal diabetes management suite derived from the MedLongMem evaluation suite. The evaluation focuses on the "hard" subset ($n = 20$ dialogues), where patient states frequently change, conflict, or are retroactively corrected. Critically, evaluation is performed at the *slot/state level*: each dialogue generates multiple `CanonicalMemory` records, yielding hundreds of individual evaluation points. All reported metrics include 95% Confidence Intervals (CI) computed via clustered bootstrap over 1000 resamples at the dialogue level, accounting for intra-dialogue correlation among state slots.

We report three primary metrics: **Unique-F1 (Strict)**, which measures the precision and recall of correctly identified unique, non-redundant state values; **Conflict-F1**, which measures the accuracy of conflict detection and branch assignment; and **QA Accuracy (State-Grounded)**, which measures the correctness of direct state queries (e.g., "What is the patient's current medication?").

6.2 Baselines

We compare against four baselines. To ensure fair comparison, all memory-augmented baselines (1, 3, 4) and our system share the identical LLM information extraction frontend (GPT-4o) to isolate the effect of memory architecture from extraction capability differences. (1) **Overwrite Update**: a naive tracker that replaces the old state with the newest extracted value; (2) **Long-Context LLM**: the full dialogue history is provided to GPT-4o in a single prompt without external memory, using a standardized system prompt optimized for state tracking; (3) **Vector Memory (RAG)**: top-K ($K = 5$) retrieval over dialogue turn embeddings (using `text-embedding-3-small`) via cosine similarity, appended to the generation prompt; and (4) **Graph-only Memory**: a temporal knowledge graph (modeled after Graphiti (Rasmussen et al., 2025)) that retrieves all 1-hop neighbor nodes and edges related to the queried entity, serialized as a textual list in the prompt, without any state arbitration mechanism.

6.3 Main Results

As shown in Table 1, our system achieves a Unique-F1 of 0.8508, more than doubling the performance of the best baseline (Overwrite Update: 0.4531). The Long-Context LLM baseline, despite having access to the full dialogue history, achieves only 0.3848 Unique-F1, confirming that attention-based context processing is insufficient for reliable long-horizon state tracking. The near-perfect Conflict-F1 of 0.9762 demonstrates that the BRANCH-CONFLICT operator, combined with the formal constraints, accurately identifies and

Table 1: Performance on the Med-LongMem longitudinal diabetes suite (hard subset, slot-level evaluation). Bootstrap 95% CIs are shown in brackets. Our system (UCM + Arbitration) substantially improves upon all baselines across all three metrics.

System	Unique-F1 (Strict)	Conflict-F1	QA Accuracy
Long-Context LLM (GPT-4o)	0.3848 [0.35, 0.42]	0.8867 [0.85, 0.92]	0.652 [0.61, 0.69]
Vector Memory (RAG)	0.3012 [0.27, 0.33]	0.6034 [0.55, 0.65]	0.415 [0.37, 0.46]
Graph-only Memory	0.0627 [0.04, 0.09]	0.3450 [0.29, 0.40]	0.288 [0.24, 0.33]
Overwrite Update	0.4531 [0.41, 0.49]	0.5012 [0.46, 0.54]	0.512 [0.47, 0.55]
Ours (UCM + Arbitration)	**0.8508** [0.82, 0.88]	**0.9762** [0.96, 0.99]	**0.895** [0.86, 0.92]

records contradictions without false positives. The Graph-only Memory baseline performs poorly on Unique-F1 (0.0627), as its retrieval mechanism returns all related graph nodes without resolving conflicts, leading to high redundancy. When queried about a medication that has changed dosages three times, the graph retrieves all three dosage nodes and their associated edges. The LLM is then forced to perform in-context arbitration, which frequently fails due to attention dilution across the long, complex graph serialization. This highlights a critical limitation of graph-based RAG in temporal domains: structural retrieval is not a substitute for state arbitration.

State-Grounded Query Accuracy. Our system achieves 89.5% QA Accuracy on state-grounded queries, compared to 65.2% for the Long-Context LLM and 41.5% for Vector Memory. This demonstrates that the confidence-thresholded escalation layer not only reduces latency (Tier 1 answers 71% of queries without any retrieval) but also improves answer quality by grounding responses in the arbitrated state rather than raw conversational fragments. Vector Memory frequently suffers from the "recency trap," retrieving the most semantically similar past conversation rather than the most temporally valid state. For example, a query about "current insulin dosage" might retrieve a highly detailed discussion from three months ago, ignoring a brief but critical dosage adjustment from last week. Our state-centric approach structurally prevents this failure mode by explicitly querying the active valid-time state.

Efficiency Analysis. Beyond accuracy, the confidence-thresholded escalation layer significantly improves query efficiency. As shown in Table 2, 71.4% of queries are resolved entirely at Tier 1 (State Check) with an average latency of just 45ms, avoiding expensive LLM generation or dense retrieval entirely. Only 3.2% of queries require full graph traversal (Tier 4). The average query latency across the suite is 312ms, compared to 1450ms for the Vector Memory baseline which performs dense retrieval and LLM synthesis for every query.

Table 2: Query Routing Distribution and Latency.

Tier	% Queries	Latency
State Check	71.4%	45ms
Structured DB	17.2%	120ms
Episodic RAG	8.2%	1450ms
Graph Traversal	3.2%	2800ms
Overall	**100%**	**312ms**

6.4 Ablation Study

Table 3 presents ablation results. Removing all arbitration operators (reverting to Overwrite Update) causes the largest performance drop, confirming that the arbitration mechanism is the primary driver of improvement. Removing only the Branch-Conflict operator while retaining Supersede maintains reasonable Unique-F1 (0.72) but collapses Conflict-F1 to 0.49, as the system can no longer represent ambiguous states. Using fixed routing (always querying Tier 3 episodic memory) achieves competitive accuracy but increases mean query latency by 3.2× compared to the confidence-thresholded escalation policy.

6.5 End-to-End Agent Evaluation

To demonstrate the impact of state arbitration on downstream agent behavior, we evaluate the full agent (UCM + Arbitration) against the Overwrite baseline on multi-turn interactions. Table 4 presents three representative cases from the evaluation suite.

In Case 1, the Overwrite baseline fails to track the temporal validity of the dosage change, giving

Table 3: Ablation results. Each row removes one component of the full system.

CONFIGURATION	UNIQUE-F1	CONFLICT-F1
FULL SYSTEM	**0.8508**	**0.9762**
W/O ARBITRATION (OVERWRITE)	0.4531	0.5012
W/O BRANCH-CONFLICT	0.7213	0.4891
FIXED ROUTING (TIER 3 ALWAYS)	0.8201	0.9650

potentially harmful advice. In Case 2, the baseline replaces the old symptom entirely, whereas the arbitration agent correctly preserves both symptoms as a concurrent state. In Case 3, the baseline is confused by retroactive reporting, while the arbitration agent correctly reconstructs the historical timeline. These cases highlight that state arbitration is not merely a memory optimization, but a prerequisite for safe, coherent agentic behavior in longitudinal settings.

6.6 Error Analysis

We performed a stratified error analysis across attribute categories (medication, vital signs, symptoms, diagnosis) and difficulty levels (easy, medium, hard) using the error analysis framework provided in the Med-LongMem suite.

The most common failure mode, accounting for 61% of Unique-F1 errors, is *false supersede*. This occurs when the system incorrectly triggers the SUPERSEDE operator because a patient uses colloquial language that superficially resembles a stop or change pattern. For example, a patient stating "I haven't been taking it regularly" might be misclassified as a definitive medication stop, terminating the valid-time window prematurely.

A secondary failure mode is *temporal anchoring failure* (22% of errors). When a patient provides a vague temporal reference, such as "a while ago" or "sometime last month," the system struggles to reliably set t_{event}. This leads to incorrect valid-time window assignments, which can subsequently cause valid states to be incorrectly flagged as conflicting or superseded when new, precisely-timestamped evidence arrives.

Finally, *implicit refinement failures* account for 12% of errors. These occur when new evidence should trigger the REFINE operator but is instead treated as a BRANCH-CONFLICT because the semantic equivalence threshold is too strict. For instance, updating "Metformin" to "Glucophage" (a brand name for Metformin) might be treated as a conflict rather than a refinement if the underlying medical ontology mapping fails.

These findings suggest that while the arbitration architecture is robust, its performance is bottlenecked by the precision of the underlying natural language understanding modules. Improving the robustness of the action pattern classifier, integrating a more sophisticated temporal expression normalizer, and expanding the medical ontology for equivalence checking are the most impactful directions for future work.

7 Discussion

7.1 Generalizability Beyond Diabetes

While our evaluation focuses on diabetes management, the state-centric paradigm is designed to generalize to any longitudinal biomedical domain where patient attributes evolve over time. Oncology treatment tracking presents a particularly compelling application: chemotherapy regimens are frequently modified based on toxicity profiles and tumor response, creating a complex sequence of superseded and branching medication states that retrieval-centric systems cannot reliably manage. Similarly, post-surgical recovery monitoring requires the agent to track wound healing progression, pain scores, and rehabilitation milestones across weeks, where early reports may be retroactively revised by subsequent clinical assessments. Mental health longitudinal assessment—tracking symptom severity, medication adherence, and psychosocial stressors over months—exemplifies the need for the BRANCH-CONFLICT operator, as patients often provide contradictory self-reports across sessions.

The key requirement for applying our framework to these domains is the specification of domain-specific formal constraints (Section 4) and the calibration of the confidence weighting hyperparameters (α, β, γ) on domain-specific data. The bi-temporal representation and the four arbitration operators are domain-agnostic; only the constraint rules and confidence priors require adaptation.

Table 4: End-to-End Agent Evaluation: Representative multi-turn cases comparing the Arbitration Agent with the Overwrite Baseline.

Dialogue Context	Overwrite Baseline	Ours (Arbitration Agent)
T1 (Jan 5): "Started Metformin 500mg." **T12 (Feb 1):** "Doc increased to 1000mg." **T15 (Feb 3):** "Should I take my 500mg pill now?"	*Yes, you can take your 500mg Metformin pill now.* (Fails to recognize the superseded dosage).	*No, your doctor increased your dosage to 1000mg on Feb 1. You should not take the 500mg pill.*
T4 (Mar 10): "I have numbness in my feet." **T8 (Mar 20):** "My vision is blurry today." **Query:** "What are my current symptoms?"	*You are currently experiencing blurry vision.* (Overwrites previous symptom).	*You are currently experiencing blurry vision (since Mar 20) and numbness in your feet (since Mar 10).*
T2 (Apr 1): "Taking 10U Insulin." **T10 (Apr 15):** "I actually stopped the Insulin three days ago." **Query:** "Was I taking Insulin on Apr 13?"	*Yes, you were taking 10U Insulin on Apr 13.* (Fails to handle retroactive reporting).	*No, you stopped taking Insulin on Apr 12 (three days before your Apr 15 report).*

7.2 Limitations and Future Directions

Our current implementation has three notable limitations. First, the action pattern classifier for detecting SUPERSEDE triggers relies on a manually curated lexical pattern set. While this achieves high precision on the evaluation suite, it may fail on colloquial or culturally specific language patterns not represented in the training data. A learned classifier, trained on a larger corpus of annotated medical dialogues, would improve robustness.

Second, the temporal expression normalizer currently handles explicit date references and relative expressions ("three days ago") but struggles with vague references ("a while ago", "recently"). Integrating a dedicated temporal information extraction model, such as those trained on the TimeML annotation scheme, would improve the accuracy of t_{event} assignment.

Third, our evaluation suite, while slot-level and statistically bootstrapped, is derived from a single disease domain (Type 2 diabetes). A multi-domain evaluation across at least three chronic disease types would provide stronger evidence for the generalizability of our approach. We release our evaluation framework and the Med-LongMem diabetes suite to facilitate future benchmarking in this area.

8 Conclusion

We have argued that longitudinal healthcare agents require a fundamental shift from retrieval-centric to state-centric memory. Our framework—comprising a bi-temporal state representation, an incremental four-operator arbitration mechanism, and a confidence-thresholded evidence escalation layer—provides a principled solution to the problem of maintaining accurate, evolving patient state models over long-horizon interactions. Evaluated on a longitudinal diabetes management suite as a representative biomedical state tracking task, our system substantially improves upon all baselines, demonstrating both the inadequacy of existing retrieval-centric approaches and the effectiveness of our state-centric alternative.

More broadly, our work illustrates that agentic AI is necessary in longitudinal biomedical settings precisely because patient states are dynamic, uncertain, and require continuous, evidence-grounded arbitration. The classical concepts of belief revision and bi-temporal databases, when operationalized for natural-language-driven agents, provide a powerful foundation for this task. We anticipate that the state-centric paradigm will generalize beyond diabetes management to other longitudinal biomedical domains, including oncology treatment tracking, post-surgical recovery monitoring, and mental health longitudinal assessment.

References

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

Jon Doyle. 1979. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272.

Shanghua Gao, Ada Fang, Yepeng Huang, and 1 others. 2024. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151.

Peter Gärdenfors. 1988. Knowledge in flux: Modeling the dynamics of epistemic states. *MIT Press*.

Zhuohan Ge, Haoyang Li, Yubo Wang, Nicole Hu, Chen Jason Zhang, and Qing Li. 2026. ClinicalAgents: Multi-agent orchestration for clinical decision making with dual-memory. *arXiv preprint arXiv:2603.26182*.

Christian S. Jensen, Michael D. Soo, and Richard T. Snodgrass. 1994. Unifying temporal data models via a conceptual model. *Information Systems*, 19(7):513–547.

Jiale Kang, Mingyu Ji, Zhiyuan Zhao, and Tao Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. ArXiv:2005.11401.

Ruoqi Li, Xin Wang, Dan Berlowitz, and 1 others. 2025. CARE-AD: A multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*.

Fang Liu and 1 others. 2025. A foundational architecture for AI agents in healthcare. *npj Digital Medicine*. PMC12629813.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. In *Advances in Neural Information Processing Systems*. ArXiv:2310.08560.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*.

Richard T. Snodgrass and Ilsoo Ahn. 1985. A taxonomy of time databases. *ACM SIGMOD Record*, 14(4):236–246.

Cheng Yang, Chuan Zhou, Yixin Xiao, Shen Dong, and Linyuan Zhuang. 2026. Graph-based agent memory: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2602.05665*.