Detecting Clinical Discrepancies in Health Coaching Agents: A Dual-Stream Memory and Reconciliation Architecture

Samuel L Pugh SAMUELPUGH@VERILY.HEALTH
Verily Health Inc.

Eric Yang ERYANG@VERILY.HEALTH
Verily Health Inc.

Alexander Muir Sutherland ASUTHERLAND@VERILY.HEALTH
Verily Health Inc.

Alessandra Breschi ABRESCHI@VERILY.HEALTH
Verily Health Inc.

Abstract

As Large Language Model (LLM) agents transition from single-session tools to persistent systems managing longitudinal healthcare journeys, their memory architectures face a critical challenge: reconciling two imperfect sources of truth. The patient's evolving self-report is current but prone to recall bias, while the Electronic Health Record (EHR) is medically validated but frequently stale. General-purpose agent memory systems optimize for coherence by overwriting older facts with the user's latest statement, a pattern that risks safety failures when applied to clinical data. We introduce a Dual-Stream Memory Architecture that strictly separates the patient narrative from the structured clinical record (FHIR), governed by a dedicated Reconciliation Engine that evaluates every extracted memory against the patient's FHIR record and classifies discrepancies by type, severity, and the specific FHIR resources involved. We evaluate this architecture on 26 patients across 675 longitudinal wellness coaching sessions, using a hybrid dataset that interleaves real provider-patient transcripts with synthetic, FHIR-grounded clinical scenarios. In isolated testing, the engine detects 84.4% of designed clinical discrepancies with 86.7% safety-critical recall. By coupling extraction and reconciliation evaluation on the same data, we directly quantify a 13.6% error cascade, tracing the degradation to clinical details lost during memory extraction from unstructured conversation rather than to downstream classification errors. These findings establish that validating patient-reported memories against clinical records is both feasible and necessary for safe deployment of longitudinal health agents.

1. Introduction

Conversational AI driven by Large Language Models (LLMs) has transformed digital health, shifting from static patient portals to highly interactive, scalable interfaces capable of continuous patient engagement, triage, and health education (Cevasco et al., 2024; Mahajan and Powell, 2025; Rashidian et al., 2025; Yang et al., 2025). However, as these systems transition from stateless, single-session query engines to persistent agents managing long-horizon longitudinal care journeys, they encounter severe architectural shortcomings. The fundamental issue lies not in language generation, but in memory architecture and epistemic uncertainty (Brender et al., 2025).

Managing patient health information requires reconciling two imperfect sources of truth: the patient and the electronic health record (EHR) (Christof and Armoundas, 2025). The patient possesses current, lived experience but is prone to recall bias (Zini and Banfi, 2021). Conversely, the EHR contains medically validated data but is frequently stale or incomplete (Gurupur et al., 2025). Current AI architectures exacerbate this dilemma. While prior work on agent memory systems has improved conversational coherence and personalization, healthcare requires strict data integrity. Standard Retrieval-Augmented Generation (RAG) systems flatten these data sources into a single vector space, treating all retrieved context as equally authoritative. Without reasoning about the decay of clinical data, these systems cannot effectively weigh a divergent patient narrative against an aging medical record (Müller et al., 2025). This architectural flaw inevitably leads to either hallucinated compliance, where the agent accepts an inaccurate self-report, or protocol rigidity, where it rigidly enforces an outdated system record.

Traditionally, healthcare systems mitigate these divergent narratives through reconciliation, an episodic, clinician-initiated workflow typically reserved for structured care transitions (Barnsteiner, 2008). However, this manual process is ill-equipped to capture the gradual drift in patient behavior and health status that emerges over weeks or months of casual conversation. Digital health agents, such as wellness coaches and chronic disease management tools, interact with patients far more frequently than traditional clinic visits. This high-frequency engagement creates a significant opportunity. Rather than waiting for the next scheduled appointment, these agents can continuously detect EHR staleness and actively reconcile discrepancies in the background of routine conversation.

To address this gap, we propose a Dual-Stream Memory Architecture for conversational health agents. Longitudinal health agents already require persistent memory to track patient goals, preferences, and context across sessions; this is not a design choice but a prerequisite for continuity of care. The core question is not whether these agents should maintain memory, but how to do so safely when patient-reported information may conflict with validated clinical data. Our architecture addresses this by strictly isolating the patient's evolving narrative from the read-only, structured clinical record (FHIR). Furthermore, we introduce a dedicated Reconciliation Engine that evaluates every new narrative memory against the patient's FHIR clinical record. Rather than silently overwriting facts, the engine actively flags conflicts, such as a patient discontinuing an active medication, and produces a discrepancy classification that traces each finding back to specific FHIR resources. This fundamentally decouples objective discrepancy detection from subjective dialogue generation.

Our core technical contributions are threefold. First, we implement the dual-stream memory and reconciliation architecture, a modular framework that maintains clinically consistent patient states by continuously validating extracted narrative memories against authoritative clinical records. Second, we introduce a novel hybrid evaluation dataset that seamlessly interleaves authentic wellness coaching transcripts with synthetic, FHIR-grounded clinical scenarios, enabling rigorous longitudinal stress-testing across 675 continuous sessions. Finally, by evaluating extraction and reconciliation both in isolation and sequentially, we provide an end-to-end error cascade analysis that directly quantifies how upstream memory extraction errors degrade downstream clinical detection, revealing critical bottlenecks in multi-stage clinical AI pipelines.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work offers three insights relevant to the broader ML-for-health community:

- **Agent memory systems for healthcare must validate patient statements against clinical records.** General-purpose memory systems treat the user's latest statement as ground truth. In healthcare, this paradigm is fundamentally unsafe; a patient reporting they stopped taking a medication should trigger a structured clinical evaluation, not a silent database update. We demonstrate that separating patient-reported memories from the clinical record, and reconciling between them at extraction time, enables reliable discrepancy detection (84.4%) while maintaining high faithfulness (4.8/5) in the stored memories.
- **Multi-stage clinical AI pipelines require component-level evaluation to identify bottlenecks.** End-to-end evaluation of chained LLM systems can mask where failures originate. By evaluating memory extraction and reconciliation both in isolation and together, we directly measure a 13.6% error cascade and trace it to the extraction stage, demonstrating that end-to-end evaluation alone would mask the true source of performance degradation.
- **Clinical reconciliation can operate as a modular background process on existing agents.** Rather than building dedicated medical interview agents, we show that EHR reconciliation can operate as a modular, ambient background process layered onto routine, non-clinical interactions (wellness coaching, chronic disease management, patient navigation). This architecture provides a blueprint for layering clinical safety guardrails onto existing, proprietary, or domain-specific conversational models.

2. Related Work

2.1. Agent Memory Systems and Benchmarks

Early cognitive frameworks for AI systems, such as Generative Agents, simulate human recall by partitioning memory into episodic, semantic, and procedural groups (Park et al., 2023). Operating system-inspired hierarchies, such as MemGPT (Letta), manage finite context windows and grant agents autonomy to retrieve information via tool calls (Packer et al., 2023). Production systems like Mem0 layer graph-based stores over vector databases to execute explicit structural operations based on semantic proximity (Chhikara et al., 2025). Temporally-aware architectures like Zep advance further by maintaining bi-temporal knowledge graphs (Rasmussen et al., 2025). These systems optimize for conversational coherence, persona alignment, and user satisfaction. When faced with contradictory information, they edit older facts to match the user's latest input. While acceptable in general domains, overwriting medical information based on casual statements degrades the factual integrity of clinical records.

This same limitation extends to how these systems are evaluated. Benchmarks like LoCoMo, LongMemEval, and MemoryBench focus heavily on an agent's explicit recall capacity rather than its ability to handle contradictions (Maharana et al., 2024; Wu et al., 2024; Ai et al., 2025). While recent benchmarks recognize conflict resolution as a critical

competency, they fundamentally assume the user narrative as a single source of truth. For instance, MemoryAgentBench (Hu et al., 2025) demonstrates that frontier models struggle with basic memory conflicts, while frameworks like BeliefShift (Myakala et al., 2026) measure an agent's susceptibility to temporal opinion drift. However, these evaluations are confined to domains where the user's internal narrative is paramount. Consequently, a critical gap remains: no existing benchmark evaluates the epistemic complexity of healthcare, where an agent must actively detect and reconcile contradictions between a subjective, continuous patient dialogue and an authoritative external database like the EHR.

2.2. Clinical Agents and the Illusion of Ground Truth

While general-domain agents prioritize user alignment, healthcare agents swing to the opposite extreme by treating the EHR as infallible. Specialized benchmarks (e.g., FHIR-AgentBench, MedAgentBench, FHIR-AgentEval) and clinical NLP tools (e.g., LLMonFHIR) evaluate an agent's ability to accurately navigate, query, and alter complex patient records (Lee et al., 2025; Jiang et al., 2025; Mokssit et al., 2026; Schmiedmayer et al., 2025; Qu et al., 2026; Alsentzer et al., 2019; Rasmy et al., 2021). By defining success purely on adherence to the provided FHIR environment, these prior works ignore the temporal staleness and fragmentation inherent in real-world EHRs, optimizing agents for a sanitized reality. This leaves them highly vulnerable to epistemic uncertainty when a patient's actual health status evolves. In contrast, our architecture explicitly rejects the assumption of EHR infallibility. We treat the clinical record not as static ground truth, but as a historical baseline that must be continuously validated against the patient's lived experience.

2.3. Conversational Medication Reconciliation

To address patient information drift, systems like AMREC actively probe patients for medication discordances (Deo et al., 2025). However, because they function as dedicated medical interviewers, they rely on structured interrogation. This makes them poorly equipped for casual, unstructured interactions where clinical information emerges organically. Similarly, LLMs have been evaluated on static medication reconciliation tasks using structured prompts rather than natural dialogue (Sridharan and Sivaramakrishnan, 2024). Rather than treating reconciliation as an explicit dialogue task, our system embeds it passively within the background memory pipeline. By extracting memories from natural conversations and flagging discrepancies against the FHIR record, our architecture can route actionable alerts without interrupting the natural conversation flow.

3. Methods

3.1. Problem Formulation

We frame clinical-narrative reconciliation as a continuous process of tracking patient state and comparing it across two distinct data streams. At dialogue turn t, given a history $H_t = (u_1, a_1, \ldots, a_{t-1}, u_t)$ of patient utterances u and agent responses a, the system evaluates extracted narrative updates against a static clinical environment. To handle the uncertainty of conflicting information, we separate the data into two independent streams. The Narrative Stream (M_N) captures patient-reported assertions, defined prior to turn t as

$M_N^{(t-1)} = \{m_1, \ldots, m_k\}$, where each memory m_i represents a single piece of patient-reported information (e.g., a health status, medication, or preference). The Clinical Stream (M_C) is a read-only FHIR R4 bundle defined as $M_C = \{e_1, \ldots, e_n\}$. Each clinical entity $e_j \in M_C$ is a tuple $e_j = (v_j, \lambda_j, \rho_j)$ containing the clinical value v_j (e.g., an active diagnosis), the resource ontology λ_j (e.g., Condition), and temporal metadata ρ_j to gauge data staleness. Instead of passively appending text to a context window, an LLM-based extraction function f_{ext} identifies what changed in each turn, producing explicit delta-operations:

$$\Delta M_N^{(t)} = f_{ext}(u_t, M_N^{(t-1)})$$

where $\Delta M_N^{(t)}$ is the subset of $\mathcal{O} \in \{\text{insert}, \text{update}, \text{delete}\}$ operations generated at turn t. By processing only these updates, the Reconciliation Engine f_{rec} compares each altered memory $m_i \in \Delta M_N^{(t)}$ against the clinical stream M_C to identify conflicts or gaps. It outputs a discrepancy characterization tuple:

$$R_i = f_{rec}(m_i, M_C) = (d, E_c, s, \sigma, \psi)$$

Here, $d \in \{0, 1\}$ indicates whether a clinically relevant discrepancy was detected (contradiction or information gap), $E_c \subseteq M_C$ identifies the FHIR resources involved, and $s \in \{\text{low}, \text{medium}, \text{high}\}$ denotes clinical significance. The boolean $\sigma \in \{0, 1\}$ indicates whether the discrepancy poses immediate patient safety risk (e.g., a contraindicated medication or unreported allergy), enabling downstream systems to prioritize urgent findings. ψ provides the model's natural language justification. This formalization strictly grounds reconciliation in traceable FHIR resources.

3.2. Dual-Stream Memory Architecture

To implement the reconciliation formalized above, we introduce a dual-stream architecture that isolates patient-reported memories from structured clinical data (Figure 1).

3.2.1. The Narrative Stream and Delta-Based Extraction

The narrative stream captures patient information via free-form content, categorical tags, and strict temporal metadata. To prevent memory fragmentation, the system separates information into two distinct types: stable facts (which receive persistent records) and evolving states (which are updated in-place to preserve their temporal trajectory). During extraction, an LLM pipeline processes the raw dialogue to generate structured `MemoryDelta` updates (Appendix D). Rather than rewriting the entire memory state from scratch, the model extracts only what has changed, outputting discrete `insert`, `update`, and `delete` operations. Duplicate prevention is rigorously enforced through prompt instructions, exact-match guards, and inline ID routing (see Appendix G for the full extraction prompt). Finally, the system maintains two memory views: an extraction view containing IDs and metadata for precise LLM editing, and a clean, categorized view for the downstream agent or practitioner.

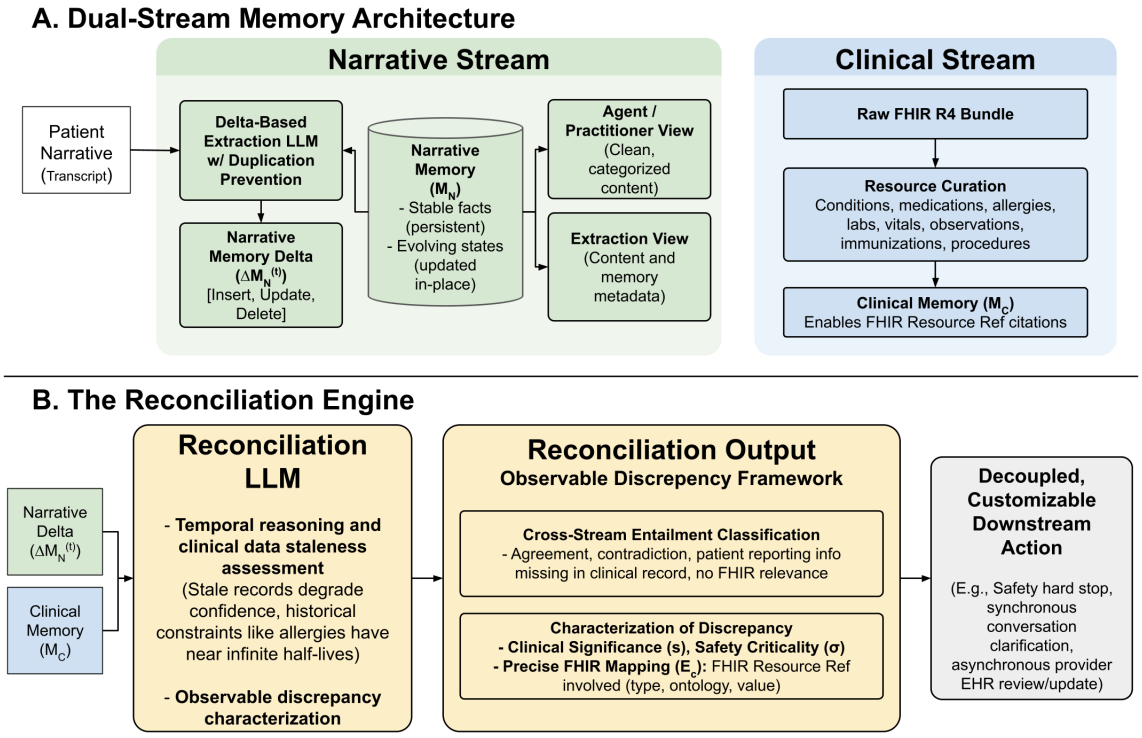

Figure 1: Dual-Stream memory architecture featuring narrative and clinical streams, a delta-based extraction system, and a reconciliation engine for discrepancy characterization and downstream action.

3.2.2. The Clinical Stream

The clinical stream (M_C) represents the patient's objective medical history, constructed from a curated FHIR R4 bundle. To optimize the context window, the clinical summary is curated to approximately 1100-1600 tokens, retaining active or resolved conditions, medications, allergies, and the most recent values for relevant observations (labs, vitals, procedures, encounters). All resources include structured inline tags formatted as `[ResourceType] display [SYSTEM:CODE]` (e.g., `[SNOMED-CT:59621000]`). This explicit tagging enables the Reconciliation Engine to output precise `FHIRResourceRef` citations, ensuring rigorous resource-level precision and recall matching during evaluation (see Appendix B for a sample clinical summary).

3.3. The Reconciliation Engine

To evaluate conflicts, the Reconciliation Engine leverages LLM-driven temporal reasoning. The model is explicitly prompted to contextualize the temporal metadata (ρ_j) of each FHIR resource. Through this qualitative reasoning process, the LLM assesses that stale records (e.g., a medication list from two years ago) warrant lower clinical confidence, whereas persistent historical constraints (e.g., allergies) retain high authority regardless of their age.

To ensure reproducible measurement, the engine translates this prompt-driven reasoning into an Observable Discrepancy Framework via structured JSON outputs. The engine classifies every narrative update into one of four states:

- **Agreement**: the patient's statement is consistent with the clinical record.
- **Contradiction**: the statement directly conflicts with documented clinical data.
- **Gap**: the patient reports clinically actionable information that is absent from the record (e.g., an undocumented medication, unreported symptom, or condition the patient reports as resolved).
- **No clinical relevance**: routine lifestyle content (step goals, preferences, daily routines) with no bearing on the FHIR record.

If an actionable discrepancy is detected, the engine characterizes it along several dimensions: clinical significance (s: low, medium, high) reflecting potential for patient harm, and a safety criticality flag (σ) indicating whether the discrepancy poses immediate patient safety risk. These dimensions are orthogonal: a moderate-severity discrepancy can be safety-critical if failing to address it could lead to acute harm. Both labels are assigned during scenario generation rather than by clinical adjudication, a limitation discussed in Section 6. The engine also outputs precise FHIR Resource Ref citations (E_c) containing the resource type, ontology, and value, ensuring that detected discrepancies are traceable to specific clinical records (see Appendices E, H, and C).

3.4. Modular Integration

The dual-stream memory and reconciliation engine is designed as a modular layer that can sit on top of any conversational agent. At each patient turn, the extraction module produces narrative deltas and the reconciliation engine evaluates them against the clinical stream, producing structured discrepancy outputs. The engine operates on extracted memories rather than raw conversation text, ensuring its inputs are self-contained semantic units that do not require conversational context for disambiguation. Importantly, the dual-stream separation governs how information is *sourced and governed* upstream: the narrative stream captures patient-reported content with different trust and write semantics than the read-only clinical record, rather than partitioning retrieval at inference time. The reconciliation engine receives both a patient memory and the full clinical summary as explicit inputs; the architectural contribution is the strict separation of provenance, not a retrieval-index design. How these outputs are acted upon (e.g., provider alerts, conversational clarification, or safety stops) is left to the deployment context and is not evaluated in this work. Here, we evaluate the extraction and reconciliation components via transcript replay, isolating their performance from both dialogue generation and downstream action handling.

4. Experimental Setup

We utilize GPT-4o for delta-based memory extraction and discrepancy classification, selected for its strong structured output capabilities and consistent adherence to schema constraints (OpenAI, 2024). All inference calls use temperature 0 with Pydantic-enforced

JSON schemas for structured output. Synthetic data generation (scenario creation, hybrid transcript generation, and ground-truth memory extraction) was performed using Gemini 2.5 Pro, chosen for its large context window which accommodates full FHIR bundles alongside conversation history in a single call. Automated evaluation judgments use GPT-4o for ground-truth memory matching (requiring precise semantic comparison) and GPT-4.1 for transcript-level quality assessment (requiring comprehension of full multi-session transcripts).

4.1. Dataset Composition

Our evaluation is grounded in authentic patient interactions by modeling a longitudinal wellness coaching agent designed to help patients set step goals, track physical activity, and discuss behavioral barriers. The foundational narrative data driving this simulation is sourced from the Virtual Coach dataset developed by the University of Illinois Chicago Natural Language Processing (UIC NLP) Lab, which contains real, de-identified text-message dialogues between patients and human health coaches (Gupta et al., 2020). Utilizing this corpus, we deterministically segmented 3,665 dialogue turns across 26 patients into 432 discrete, multi-turn wellness coaching sessions using temporal heuristics: messages within a 4-hour window were grouped into the same session, with a maximum session span of 48 hours (average 8.5 turns per session). While highly realistic, these authentic transcripts primarily focus on lifestyle habits and behavioral motivation, naturally lacking the clinical contradictions required to rigorously stress-test a medical reconciliation engine.

Because comprehensive system evaluation demands a robust clinical environment, we inferred clinical demographic profiles from the conversational dataset and passed these parameters into Synthea to generate 50 to 100 candidate FHIR R4 bundles per patient (Walonoski et al., 2018). These candidates were evaluated using a multi-factor scoring model prioritizing condition and medication overlap, resulting in 26 highly plausible, temporally aligned FHIR bundles, one per patient. Each selected bundle represents a comprehensive longitudinal clinical history, averaging 26 unique conditions (e.g., hypertension, prediabetes, chronic pain), 7 distinct medications, 47 documented procedures, and 3 immunization records per patient, consistent with the multi-morbidity profiles typical of the UIC coaching population. Synthea parameters were configured per patient using inferred age, gender, and target condition modules derived from the conversational data. Using each patient's FHIR record and UIC transcript as context, we generated 243 synthetic reconciliation scenarios (9.3 per patient on average), of which 83 (34%) are designated safety-critical. The scenarios reference 390 individual FHIR resources across seven resource types (see Table 3 for the per-type breakdown), with Conditions (123 references) and MedicationRequests (109) being the most frequently tested. Severity is distributed across low (119 scenarios), medium (71), and high (53). Crucially, all scenarios maintain the coaching scope constraint: clinical information is introduced not via explicit medical questionnaires, but organically as patients discuss physical barriers or context for their activity goals.

4.2. Hybrid-Real Simulation Strategy

Maintaining conversational cohesiveness is paramount when evaluating a longitudinal memory system. Rather than testing the reconciliation engine on isolated clinical vignettes, we

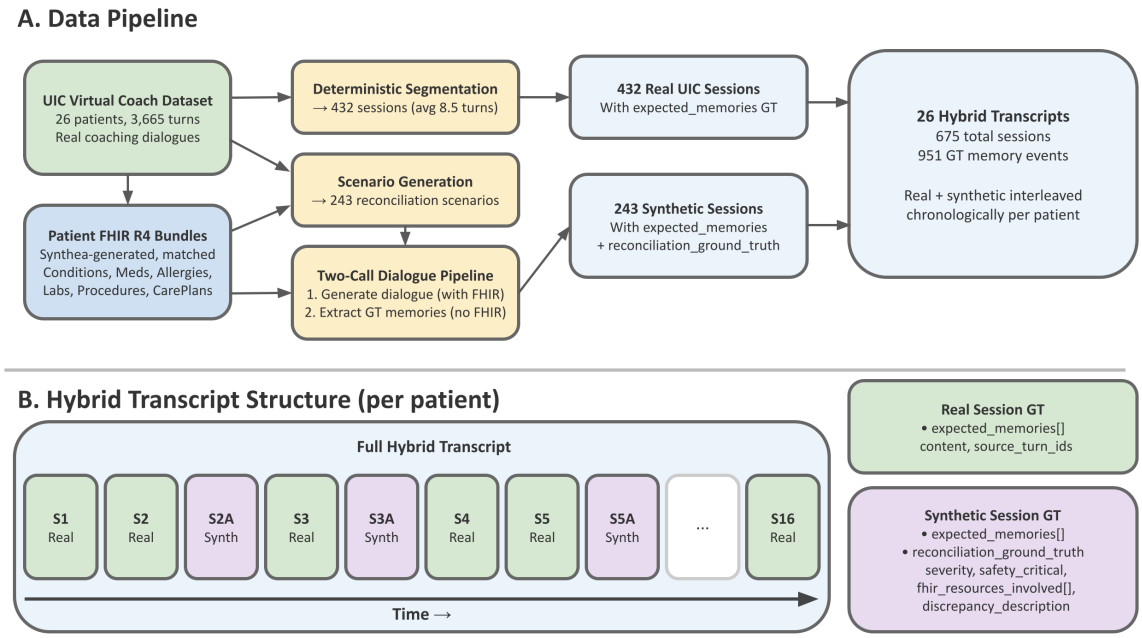

Figure 2: Hybrid transcript construction pipeline. (A) Real UIC coaching transcripts are deterministically segmented into 432 sessions, while patient-matched FHIR R4 bundles are used to generate 243 synthetic reconciliation scenarios via a two-call pipeline that separates dialogue generation (with FHIR access) from ground-truth memory extraction (transcript only). (B) Real and synthetic sessions are interleaved chronologically per patient, producing 26 hybrid transcripts with 675 total sessions and 951 ground-truth memory events. Real sessions carry expected memory annotations; synthetic sessions additionally carry reconciliation ground truth (severity, safety flag, FHIR resource references).

sought to simulate a realistic coaching trajectory where clinical issues emerge naturally amidst baseline wellness discussions. We achieved this by interleaving the full conversational dialogues derived from the synthetic clinical scenarios directly into the chronological sequence of the real UIC coaching sessions (See Figure 2). The resulting dataset comprises 26 unified hybrid transcripts totaling 675 sessions (432 real UIC sessions plus 243 synthetic sessions) and yielding 951 ground-truth memory events. Integrating authentic conversational dynamics with controlled clinical injections provides the continuous longitudinal evaluation data necessary for rigorous end-to-end testing.

Generating these synthetic scenarios required a two-call pipeline (Figure 2): dialogue generation with full FHIR access, followed by ground-truth memory extraction from the transcript only (deliberately isolated from FHIR). This extraction was applied to both synthetic and real UIC sessions, ensuring unified ground-truth memory annotations across all 675 sessions.

4.3. Multi-Dimensional Evaluation Framework

Evaluating complex, multi-stage pipelines requires isolating individual component performance before assessing the system holistically. Replaying fixed conversational transcripts rather than simulating live, dynamic agent responses guarantees reproducibility and insulates our metrics from downstream dialogue generation variance. We therefore structure our evaluation across three distinct dimensions, allowing us to pinpoint exact failure modes within the architecture before measuring the compounding effects of system-wide integration.

Dimension 1: Memory Extraction Quality. The first dimension measures the accuracy of the delta-based memory extraction pipeline across the full hybrid transcript corpus (675 sessions, 2,296 patient turns). The evaluation targets 951 ground-truth memory events: 522 from real UIC sessions (1.2 per session on average) and 429 from synthetic reconciliation sessions (1.8 per session). For each ground-truth event, an LLM judge evaluates whether the expected information is captured in the system's memory state at the corresponding point in the conversation. The judge considers information spread across multiple memories, classifying each event as MATCH (fully captured), PARTIAL (gist present but key details missing), or NO_MATCH (not captured). We report *recall* (match + partial / total) and *strict recall* (match only / total) as primary metrics. A separate transcript-level LLM judge assesses the final memory state against the complete conversation for *faithfulness* (absence of hallucinated content) and *deduplication* (absence of redundant memories), each scored 1–5 (see Appendix I for both judge prompts).

Dimension 2: Isolated Reconciliation. The second dimension assesses the reconciliation engine in isolation, independent of upstream extraction errors. Ground-truth narrative memories are fed directly into the classification module, and performance is measured against the 243 synthetic reconciliation scenarios. Each scenario carries a ground-truth label specifying the expected discrepancy, its clinical severity (low, medium, or high), a safety-critical flag, and the specific FHIR resources involved. We evaluate five metrics: (1) *binary detection*: whether the engine flagged any discrepancy in the session; (2) *resource-informed detection*: whether the engine flagged a discrepancy and cited at least one of the expected FHIR resources, measured via exact `(resource_type, code_system, code_value)` tuple matching; (3) *safety recall*: the fraction of safety-critical scenarios detected; (4) *severity accuracy*: both exact match and within-one-level agreement with ground-truth severity; and (5) *resource recall and precision*: the overlap between engine-cited and ground-truth FHIR resource codes. To characterize the engine's behavior on naturalistic data, we also process memories extracted from the 432 real UIC sessions, which contain no designed discrepancies, and report the distribution of classification types.

Dimension 3: End-to-End Pipeline and Error Cascade. The final dimension evaluates the complete pipeline by processing the hybrid transcripts through extraction and then into reconciliation. The reconciliation stage consumes the *exact same* extracted memories produced during Dimension 1, coupling the two stages and ensuring the error cascade analysis reflects the actual impact of extraction quality on downstream detection. Comparing detection accuracy between Dimensions 2 and 3 quantifies how upstream extraction failures (e.g., lost clinical details, over-consolidated memories) degrade downstream discrepancy detection.

Table 1: Memory extraction quality across 26 patients (675 hybrid sessions, 951 ground-truth events). Recall measures the fraction of GT events captured fully or partially in the memory state. Faithfulness and deduplication are scored 1–5 by an LLM transcript judge.

Metric	**Mean [95% CI]**	**Range**
GT Recall (match + partial)	0.708 [0.674, 0.740]	0.49–0.88
Strict Recall (match only)	0.262 [0.221, 0.305]	0.07–0.53
Faithfulness (/5)	4.8 [4.62, 4.92]	4–5
Deduplication (/5)	3.6 [3.35, 3.92]	3–5
Patient turns processed	2,296	
Final memories extracted	408	

5. Results

We evaluated 26 patients across 675 sessions (432 real UIC + 243 synthetic reconciliation), comprising 2,296 patient turns and 951 ground-truth memory events.

5.1. Dimension 1: Memory Extraction

Table 1 summarizes extraction performance. The system captures 70.8% of ground-truth events (25.2% full match, 44.7% partial, 30.1% no match). Partial matches typically reflect the consolidation architecture working as designed: the system maintains the current state of evolving information (e.g., the latest step goal) rather than preserving every historical detail. Faithfulness scores are high (4.8/5), indicating that the system rarely fabricates information not present in the conversation. Deduplication (3.6/5) is the weakest dimension, with goal-tracking conversations prone to redundant memory entries when patients express similar intentions across sessions. Although deduplication is the lowest-scoring extraction dimension, the cascade loss identified in Section 5.3 primarily stems from a distinct failure mode: clinical detail lost during memory consolidation (e.g., a specific drug name absorbed into a broader lifestyle memory), rather than from redundancy itself. Improving extraction for downstream reconciliation therefore requires preserving clinical specificity during updates, not solely reducing duplication.

5.1.1. Human Expert Validation of Extracted Memories

Two domain experts independently reviewed transcripts and rated all 104 GT memory events across 4 patients (80 sessions) for accuracy and clinical relevance using the rubric in Appendix F. Neither annotator rated any event as Inaccurate, and both agreed on at least Partially Accurate for 100% of events and at least Useful for 99% of events. Annotators showed 100% within-one-level agreement on both dimensions, with Gwet's AC1 (Gwet, 2008) indicating substantial agreement (0.80 for accuracy, 0.67 for relevance). Review of annotator-flagged missing information identified only 6 instances across 80 sessions where

Table 2: Reconciliation performance on 243 synthetic scenarios using ground-truth memories (Dimension 2) versus extracted memories from the full pipeline (Dimension 3). The delta column quantifies the error cascade from upstream extraction. 95% bootstrap confidence intervals in brackets.

Metric	**Isolated (GT Memories)**	**Full Pipeline**	**Δ**
Detection rate	84.4% [79.8, 88.9]	70.8% [65.0, 76.5]	−13.6%
Resource-informed detection	76.5% [71.2, 81.9]	62.6% [56.4, 68.7]	−14.0%
Safety recall	86.7% [79.5, 94.0]	75.9% [66.3, 84.3]	−10.8%
Severity within-1	81.5% [76.5, 86.4]	67.1% [61.3, 72.8]	−14.4%
Severity exact match	28.0% [22.6, 33.7]	25.5% [20.2, 30.9]	−2.5%
Resource recall (mean)	65.3% [60.0, 70.4]	51.4% [45.9, 56.9]	−13.9%
Resource precision (mean)	71.2% [66.3, 76.1]	62.1% [56.5, 67.7]	−9.1%

information within the system's intended scope was not extracted. These results validate the quality of the GT memory events used throughout our evaluation.

5.2. Dimension 2: Isolated Reconciliation

Table 2 presents reconciliation accuracy. In isolated testing (Dimension 2), the engine detects 84.4% of designed discrepancies with 86.7% safety recall. We report two detection metrics: *binary detection* (did the engine flag any discrepancy in the session?) and *resource-informed detection* (did it cite at least one of the expected FHIR resources?). The 7.9% gap between these metrics (84.4% vs. 76.5%) reflects cases where the engine identified the correct clinical concern but cited a related rather than identical FHIR resource, for example citing the underlying condition rather than the specific medication. Review of these 19 cases confirmed that each identified the correct clinical concern but cited a related resource (e.g., a broader condition code rather than a specific procedure, or a diagnosis rather than the associated care plan). Future evaluations could consider partial-credit ontology matching (e.g., recognizing a condition code as clinically related to the expected medication) rather than requiring exact code-system tuple matches, which would better reflect clinical equivalence. Of the 205 detected discrepancies in Dimension 2, 29 (14.1%) were classified as contradictions and 176 (85.9%) as information gaps, consistent with coaching conversations surfacing undocumented patient information more frequently than direct conflicts with the clinical record.

Severity classification achieves 81.5% within-one-level accuracy but only 28.0% exact match. The engine tends to underestimate severity for complex, multi-factor scenarios (e.g., a patient planning high-intensity exercise with concurrent hypertension and a recent procedure) while occasionally overestimating low-severity information gaps. Detection rates were consistent across severity levels (84.0% low, 83.1% medium, 86.8% high), with all levels experiencing a similar 13–14% cascade loss through the full pipeline.

5.2.1. Clinical Adjudication of Severity and Safety Labels

To validate the LLM-generated labels, a domain expert independently reviewed a stratified random sample of 30 scenarios (see Appendix J for details). The clinician agreed with safety-criticality labels in 80.0% of cases and with severity within one level in 90.0% of cases, with all disagreements unidirectional: the clinician consistently rated higher severity and more scenarios as safety-critical than the LLM. The paper's primary metrics (binary detection, resource-informed detection, and the error cascade) are independent of these labels; for downstream action, a coarser binary routing scheme (safety-critical vs. not safety-critical) is likely more robust than fine-grained severity classification.

5.3. Dimension 3: Error Cascade

The coupled pipeline reveals a 13.6% absolute reduction in detection rate from Dimension 2 to Dimension 3 (84.4% $\rightarrow$ 70.8%). Of the 39 detections lost through the cascade, 54% involved low-severity scenarios, 28% medium, and 18% high-severity. Ten safety-critical scenarios were lost, in each case because the memory extraction system failed to capture the relevant clinical detail, for example consolidating a medication adherence statement into a broader lifestyle memory that lost the specific drug reference. Six scenarios were gained in Dimension 3 (detected in the full pipeline but missed in isolation), in cases where the extraction system produced memories that framed the clinical content differently than the ground truth, occasionally triggering detection where the original GT phrasing did not. The 13.6% detection degradation is statistically significant (bootstrap 95% CI for the paired difference: [8.6%, 18.9%]). Table A.1 in Appendix A illustrates representative success and failure cases from the pipeline.

5.4. Effect of Conversation Length

Extraction recall shows a moderate negative correlation with the number of patient turns ($r = -0.46$, $p = 0.018$) and with the final memory count ($r = -0.48$, $p = 0.013$). This correlation accounts for the substantial patient-level variance in recall (range 0.49–0.88 in Table 1): patients with longer transcripts and more accumulated memories experience greater recall degradation as the consolidation system merges and updates memories over many sessions. In contrast, reconciliation detection rate shows no significant correlation with conversation length (full pipeline detection rate vs. patient turns: $r = 0.12$, $p = 0.56$; vs. final memory count: $r = 0.02$, $p = 0.93$), confirming that the engine performs consistently regardless of accumulated history. This asymmetry suggests that memory extraction and management, not reconciliation capacity, is the primary bottleneck for longitudinal deployment.

To assess specificity, we ran the engine on the 432 real UIC sessions, which contain no designed discrepancies. Of 1,124 reconciliation results, 83.3% were classified as no clinical relevance, 16.6% as information gaps representing real clinical content patients mentioned in conversation (e.g., medication changes, symptom reports), and zero contradictions, indicating the engine does not fabricate conflicts where none exist.

Table 3: Detection rate by FHIR resource type. All seven resource types in the ground truth are covered, with AllergyIntolerance and Immunization achieving perfect detection in isolation.

Resource Type	Scenarios	Isolated	Full Pipeline
AllergyIntolerance	6	100.0%	66.7%
CarePlan	60	66.7%	58.3%
Condition	121	86.0%	74.4%
Immunization	25	100.0%	92.0%
MedicationRequest	92	89.1%	77.2%
Observation	28	82.1%	57.1%
Procedure	32	75.0%	65.6%

Table 4: Model comparison across all dimensions. Metrics defined in Tables 1 and 2.

	Dim. 1: Extraction				Dim. 2: Isolated			Dim. 3: Pipeline		
	Recall	Strict	Faith.	Dedup.	Det.	Safety	Res.-inf.	Det.	Safety	Res.-inf.
GPT-4o	0.708	0.262	4.8	3.6	84.4%	86.7%	76.5%	70.8%	75.9%	62.6%
GPT-5.4	0.674	0.312	5.0	4.7	73.3%	80.7%	69.5%	56.4%	66.3%	49.8%

5.5. FHIR Resource Type Coverage

Table 3 shows detection rates across all seven FHIR resource types. Discrete, fact-based resources (AllergyIntolerance and Immunization at 100%, MedicationRequest at 89.1%) achieve the highest isolated detection rates. Contextual resources (CarePlan at 66.7%, Procedure at 75.0%) are harder to detect, reflecting the difficulty of identifying conflicts with care plan recommendations and laboratory values in unstructured coaching dialogue.

5.6. Model Sensitivity

A natural question is whether the architecture's performance is driven by the specific LLM or by the architectural design itself. To investigate, we repeated the full evaluation using GPT-5.4, a more recent model, in place of GPT-4o for both memory extraction and discrepancy classification. Both models were run with temperature 0 and structured output via Pydantic JSON schemas; GPT-5.4 used the default reasoning configuration (`reasoning_effort`: none). Systematically varying reasoning budget is an unexplored axis that may affect the precision-recall tradeoff described below. Table 4 presents the comparison across all three dimensions.

GPT-4o achieves higher detection (84.4% vs. 73.3%), safety recall (86.7% vs. 80.7%), and resource-informed detection (76.5% vs. 69.5%), while GPT-5.4 produces higher faithfulness (5.0 vs. 4.8) and better deduplication (4.7 vs. 3.6). This pattern suggests a precision-recall tradeoff: GPT-5.4 is more conservative, defaulting to agreement or no clinical relevance where GPT-4o commits to flagging a potential discrepancy. For clinical surveillance where missed discrepancies carry higher risk than false alarms, GPT-4o's behavior is preferable.

Resource recall is comparable between models (65.3% vs. 66.0%), indicating that when either model detects a discrepancy, it identifies the relevant FHIR resources with similar accuracy. These results confirm that the architecture's contribution is the framework and evaluation methodology, not any particular model configuration.

To validate the stability of LLM-as-judge evaluations across model families, we re-ran the GT event matching judge (Appendix I) on a subset of 272 events (28.6%) using Claude Sonnet 4 with identical inputs. Claude agreed with GPT-4o verdicts in 83.5% of cases (227/272), with disagreements concentrated in adjacent categories (MATCH vs. PARTIAL) rather than opposite ends, indicating that the recall metrics are robust across model families.

6. Discussion

6.1. Principal Findings

Our work demonstrates that conversational agents can actively reconcile evolving patient narratives with clinical records as a background process. While existing memory systems optimize for coherence by treating the user's latest statement as ground truth, the dual-stream architecture maintains the integrity of both sources by validating extracted memories against a FHIR-grounded clinical profile. The engine detects 84.4% of designed discrepancies in isolation with 86.7% safety-critical recall and high faithfulness (4.8/5), showing that automated discrepancy detection is feasible without introducing hallucinated clinical content.

By evaluating extraction and reconciliation both in isolation and end-to-end, we directly quantified a 13.6% error cascade from memory extraction to discrepancy detection. This analysis reveals that the degradation stems from upstream extraction losing clinical details during memory consolidation, not from downstream reasoning errors. Extraction recall degrades with conversation length ($r = -0.46$, $p = 0.018$) while reconciliation detection does not ($r = 0.12$, $p = 0.56$), directly identifying extraction as the primary bottleneck for longitudinal deployment. These findings highlight that multi-stage clinical AI systems require component-level evaluation to identify where improvements matter most.

6.2. Clinical Implications

In current practice, medication and record reconciliation is an episodic, clinician-driven task, leaving informational gaps when patients adjust behaviors or experience adverse events between formal visits. By demonstrating that clinical validation can operate continuously in the background of unstructured wellness coaching conversations, this work shows the feasibility of ambient clinical surveillance through routine patient-agent interactions.

Because the reconciliation engine isolates detection from dialogue generation, it avoids interrupting conversational flow to act as a medical interrogator. Instead, it produces structured, severity-graded outputs mapped to specific FHIR resources, enabling flexible downstream actions: safety-critical findings can be routed directly to care teams, moderate discrepancies can prompt conversational clarification, and low-severity information gaps can be logged for the next scheduled visit. This modularity allows the reconciliation layer to be added to existing digital health applications (wellness coaching, chronic care management, patient navigation) without modifying the primary conversational experience, though de-

termining how best to act on detected discrepancies (e.g., when to alert a clinician versus prompt the patient for clarification) remains an open question for future deployment studies. While severity exact-match accuracy is low (28.0%), the within-one-level accuracy of 81.5% means the engine rarely confuses low-severity with high-severity scenarios. In practice, a coarser binary routing scheme (safety-critical vs. not safety-critical), directly measured by safety recall (86.7%), may be more robust than fine-grained three-level severity routing.

6.3. Limitations and Future Directions

To evaluate the pipeline across hundreds of longitudinal sessions, we used Synthea-generated FHIR bundles matched to real patients via fuzzy scoring, and LLM-generated ground-truth labels for severity and safety criticality. Clinical adjudication of a 30-scenario subset confirmed 80% safety-criticality agreement and 90% severity within-one-level agreement, though with systematic LLM underestimation of severity (Section 5.2.1). While binary detection and resource identification are robust to label quality, severity and safety recall metrics should be interpreted with this caveat. The synthetic conversations may not fully capture the ambiguity of real patient speech, and all 243 scenarios are designed to be detectable; future evaluations should include adversarial scenarios. The architecture currently operates as detection-only with a read-only clinical stream; the delta-based processing (Section 3.1) prevents re-flagging when no new information arises, but a complete write-back protocol remains future work.

Extraction recall degrades with conversation length, and scaling strategies such as semantic retrieval during consolidation or hierarchical memory compression may mitigate this but have not been evaluated. Detection rates for contextual FHIR resources (CarePlan at 66.7%, Procedure at 75.0%) lag behind discrete resources (AllergyIntolerance and Immunization at 100%), reflecting the difficulty of identifying care plan adherence conflicts in unstructured dialogue. Real EHRs would further introduce free-text notes, inconsistent coding, and missing temporal metadata; binary detection does not depend on structured ontology codes, but resource-informed detection would be affected. A prospective deployment study with real patients, real EHR data, and clinician-adjudicated labels would test whether the patterns observed here generalize to clinical practice.

Acknowledgments

We thank Theresa McKnight and Kimerly Caswell for their expert clinical annotations.

References

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. MemoryBench: A benchmark for memory and continual learning in LLM systems. December 2025.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B A McDermott. Publicly available clinical BERT embeddings. April 2019.

J. H. Barnsteiner. Medication reconciliation. In R. G. Hughes, editor, *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, chapter 38. Agency for Healthcare

Research and Quality (US), Rockville (MD), apr 2008. URL https://www.ncbi.nlm.nih.gov/books/NBK2648/.

Teva D Brender, Leo A Celi, and Julien M Cobert. Clinical notes as narratives: Implications for large language models in healthcare. *J. Gen. Intern. Med.*, 40(3):687–689, February 2025.

Kevin E Cevasco, Rachel E Morrison Brown, Rediet Woldeselassie, and Seth Kaplan. Patient engagement with conversational agents in health applications 2016-2022: A systematic review and meta-analysis. *J. Med. Syst.*, 48(1):40, April 2024.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. April 2025.

Michael Christof and Antonis A Armoundas. Implications of integrating large language models into clinical decision making. *Commun. Med. (Lond.)*, 5(1):490, November 2025.

Rahul C Deo, Shinichi Goto, Tara Jain, Sarah Meier, and Rahul Patel. A conversational artificial intelligence agent for medication reconciliation and review. June 2025.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.30.

Varadraj Gurupur, Sahar Hooshmand, Deepa Fernandes Prabhu, Elizabeth Trader, and Sanket Salvi. Incompleteness of electronic health records: An impending process problem within healthcare. *Healthcare (Basel)*, 13(22):2900, November 2025.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. July 2025. URL https://arxiv.org/abs/2507.05257. Accepted at ICLR 2026.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500144.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering. November 2025.

Arjun Mahajan and Dylan Powell. Transforming healthcare delivery with conversational AI platforms. *NPJ Digit. Med.*, 8(1):581, September 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. February 2024.

Youssef Mokssit, Kamalakkannan Ravi, Mengshu Nie, Junyoung Kim, and Cong Liu. FHIR-AgentEval: A modular sandbox for benchmarking clinical LLM agents with an evaluation of memory-augmented configurations. February 2026.

Leopold Müller, Joshua Holstein, Sarah Bause, Gerhard Satzger, and Niklas Kühl. Data quality challenges in Retrieval-Augmented generation. October 2025.

Praveen Kumar Myakala et al. Beliefshift: Benchmarking temporal belief consistency and opinion drift in LLM agents. March 2026.

OpenAI. GPT-4o system card. October 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. MemGPT: Towards LLMs as operating systems. October 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Yixiang Qu, Yifan Dai, Shilin Yu, Pradham Tanikella, Malvika Pillai, Walter Chen, Jialiu Xie, Yishan Ren, Duan Wang, Yikai Wang, Sid Sheth, Guanting Chen, Yufeng Liu, Travis Schrank, Trevor Hackman, Didong Li, and Di Wu. PrecLLM: A privacy-preserving framework for efficient clinical annotation extraction from unstructured EHRs using small-scale LLMs. March 2026.

Sina Rashidian, Nan Li, Jonathan Amar, Jong Ha Lee, Sam Pugh, Eric Yang, Geoff Masterson, Myoung Cha, Yugang Jia, and Akhil Vaid. AI agents for conversational patient triage: Preliminary simulation-based evaluation with real-world EHR data. 2025.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. January 2025. URL https://arxiv.org/abs/2501.13956.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.*, 4(1):86, May 2021.

Paul Schmiedmayer, Adrit Rao, Philipp Zagar, Lauren Aalami, Vishnu Ravi, Aydin Zahedivash, Dong-Han Yao, Arash Fereydooni, and Oliver Aalami. LLMonFHIR: A physician-validated, large language model-based mobile application for querying patient electronic health data. *JACC Adv.*, 4(6 Pt 1):101780, June 2025.

Kannan Sridharan and Gowri Sivaramakrishnan. Unlocking the potential of advanced large language models in medication review and reconciliation. *Explor. Res. Clin. Soc. Pharm.*, 15:100492, September 2024. doi: 10.1016/j.rcsop.2024.100492.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.*, 25(3):230–238, March 2018.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on Long-Term interactive memory. October 2024.

Eric Yang, Tomas Garcia, Hannah Williams, Bhawesh Kumar, Martin Ramé, Eileen Rivera, Yiran Ma, Jonathan Amar, Caricia Catalani, and Yugang Jia. From barriers to tactics: Development study of behavioral science-informed agentic workflow for personalized nutrition coaching. *JMIR Form. Res.*, 9(v9i8e75421):e75421, August 2025.

Michela Luciana Luisa Zini and Giuseppe Banfi. A narrative literature review of bias in collecting patient reported outcomes measures (PROMs). *Int. J. Environ. Res. Public Health*, 18(23):12445, November 2021.

Appendix A. Qualitative Examples

Table A.1: Qualitative examples of end-to-end pipeline behavior. The success case shows a medication discrepancy correctly detected with precise FHIR citation. The failure case shows extraction losing safety-critical details, causing the engine to miss a documented allergy conflict.

	Success: Medication Change	**Failure: Allergy Conflict**
Patient utterance	"I stopped taking that lisinopril a few days ago because I think it was making me dizzy."	"I saw some peanut butter ones that looked good for a quick boost. I'm not allergic to anything major, so it should be fine."
GT memory	"Patient stopped taking lisinopril because they believe it was causing dizziness."	"Patient is considering eating a peanut butter protein bar and reports having no major allergies."
Extracted memory	"Patient stopped taking lisinopril a few days ago due to dizziness (reported Aug 21, 2019)."	"The patient is considering taking a snack, like a protein bar, during walks to help with energy levels."
FHIR record	`MedicationRequest`: lisinopril 10 MG (active)	`AllergyIntolerance`: Allergy to peanuts (active)
Engine result	`contradiction`, severity: high	`no_fhir` (routine snack preference)
Analysis	Extraction preserved the clinical detail. Engine correctly identified the conflict with the active prescription.	Extraction dropped both "peanut butter" and the allergy denial ("not allergic to anything major"), reducing the memory to a generic snack preference. Without those details, the engine could not identify the conflict with the documented peanut allergy.

Appendix B. Sample Clinical Summary

The following is an excerpt of the curated clinical summary provided to the reconciliation engine for one patient. Each resource includes inline code tags (e.g., `[SNOMED-CT:59621000]`) enabling the LLM to cite specific FHIR resources in its structured output. The full summary for this patient contains 33 conditions, 9 medications, 45 procedures, and 6 care plan activities across approximately 1,100 tokens. The excerpt below shows only a representative subset.

```
CONDITIONS:
  - [Condition] Essential hypertension (disorder)
    (active) [onset: 1996-02-20] [SNOMED-CT:59621000]
  - [Condition] Prediabetes
```

```
    [onset: 1999-03-09] [SNOMED-CT:15777000]
  - [Condition] Anemia (disorder)
    [onset: 2000-03-14] [SNOMED-CT:271737000]
  - [Condition] Chronic sinusitis (disorder)
    [onset: 2009-09-04] [SNOMED-CT:40055000]
  - [Condition] Body mass index 30+ - obesity (finding)
    [onset: 2015-06-09] [SNOMED-CT:162864005]
  - [Condition] Ischemic heart disease (disorder)
    [onset: 2018-06-26] [SNOMED-CT:414545008]

MEDICATIONS:
  - [MedicationRequest] lisinopril 10 MG Oral Tablet
    (active) [authored: 2019-07-02] [RxNorm:314076]
  - [MedicationRequest] Clopidogrel 75 MG Oral Tablet
    (active) [authored: 2018-07-03] [RxNorm:309362]
  - [MedicationRequest] Simvastatin 20 MG Oral Tablet
    (active) [authored: 2018-07-03] [RxNorm:312961]
  - [MedicationRequest] Amoxicillin 250 MG Oral Capsule
    (stopped) [authored: 2010-04-19] [RxNorm:308182]

CARE PLANS:
  - [CarePlan] Prediabetes: Diabetic diet
    [SNOMED-CT:160670007]
  - [CarePlan] Prediabetes: Exercise therapy
    [SNOMED-CT:229065009]
  - [CarePlan] Essential hypertension: Prescribed
    activity/exercise education [SNOMED-CT:386463000]

KEY OBSERVATIONS (most recent):
  - [Observation] Body mass index (BMI) [Ratio]
    = 30.44 kg/m2 [2019-07-02] [LOINC:39156-5]
  - [Observation] Blood Pressure panel
    = Systolic: 107.0 mmHg / Diastolic: 73.0 mmHg
    [2019-07-02] [LOINC:85354-9]
  - [Observation] Hemoglobin A1c/Hemoglobin.total
    = 5.89 % [2019-07-02] [LOINC:4548-4]
```

Appendix C. Sample Reconciliation Result

The following is an actual engine output for a patient who mentions expecting an Amoxicillin prescription while claiming to have no allergies. The engine identifies the relevant FHIR resources (chronic sinusitis condition and stopped Amoxicillin prescription) and flags the clinical gap.

```
Input memory: "Patient reports a recurring sinus issue
  is acting up, has a doctor's appointment for it
  tomorrow, and expects to be prescribed Amoxicillin,
  noting they have no allergies."
Category: health
```

```
Output:
{
  "contradiction_type": "gap_patient",
  "confidence": 0.95,
  "clinical_severity": "medium",
  "justification": "The patient's memory indicates a
    recurring sinus issue and expectation of being
    prescribed Amoxicillin, which is not currently
    active in the clinical record. The clinical record
    shows a history of chronic sinusitis but does not
    reflect the current recurrence or the anticipated
    prescription of Amoxicillin. This discrepancy is
    clinically actionable as it involves a potential
    change in medication management.",
  "fhir_resources_considered": [
    {
      "resource_type": "Condition",
      "code_system": "SNOMED-CT",
      "code_value": "40055000",
      "display": "Chronic sinusitis (disorder)"
    },
    {
      "resource_type": "MedicationRequest",
      "code_system": "RxNorm",
      "code_value": "308182",
      "display": "Amoxicillin 250 MG Oral Capsule
        (stopped)"
    }
  ]
}
```

Appendix D. MemoryDelta Schema

The extraction LLM returns a structured output conforming to the following schema:

```
class MemoryInsert:
    content: str     # Memory content to store
    category: str    # preference|health|lifestyle|
                     # medication|fact

class MemoryUpdate:
    memory_id: str   # ID of existing memory to update
    new_content: str # Replacement content
    category: str

class MemoryDelete:
    memory_id: str   # ID of existing memory to remove
    justification: str

class MemoryDelta:
    inserts: List[MemoryInsert]
```

```
    updates: List[MemoryUpdate]
    deletes: List[MemoryDelete]
```

Appendix E. Reconciliation Output Schema

```
class FHIRResourceRef:
    resource_type: str  # e.g., "MedicationRequest"
    code_system: str    # e.g., "RxNorm"
    code_value: str     # e.g., "314076"
    display: str        # e.g., "lisinopril 10 MG"

class ReconciliationResult:
    contradiction_type: str  # agreement|contradiction|
                             # gap_patient|no_fhir
    confidence: float        # 0.0-1.0
    justification: str
    clinical_severity: str   # low|medium|high
    fhir_resources_considered: List[FHIRResourceRef]
```

Appendix F. Human Annotation Instructions

Two digital health practitioners with clinical backgrounds independently annotated extracted memory events for a subset of patients. The following instructions were provided verbatim:

```
MEMORY EXTRACTION ANNOTATION

Thank you for participating in this evaluation. You are
reviewing an AI system that extracts "memory events" from
health coaching conversations | facts, preferences, and
events that a coach should remember about a patient for
future sessions.

WHAT YOU'LL SEE

Each patient has their own tab. Within each tab:

  - Session transcripts: These come from real text-message
    coaching conversations conducted over several weeks.
    The original data is one long thread of messages; we
    segmented it into sessions based on gaps in the
    conversation. Sessions labeled "UIC" are these real
    coaching conversations. Sessions labeled "Clinical"
    are simulated clinical check-in sessions we created
    to test the system's ability to capture health-related
    information (e.g., medication changes, symptoms).
    Read all sessions in order, top to bottom.

  - Memory events: After each session's transcript, you'll
    see a list of facts the AI extracted. Your job is to
```

```
    rate each one.

  - Missing memories: After each session's events, there's
    a prompt to note anything the AI missed.

HOW TO ANNOTATE

For each memory event (white rows with an Event ID),
select a rating from the dropdown in two columns:

ACCURACY | Does this memory event correctly reflect what
the patient communicated?

  Accurate:
    The event faithfully and completely captures what the
    patient said. The meaning is preserved without
    distortion, exaggeration, or omission of key context.

  Partially Accurate:
    The core idea is correct, but there is a meaningful
    issue:
    - Important qualifying detail is missing (e.g.,
      "takes medication" but omits "only on weekdays")
    - Timeframe or frequency is slightly off
    - Phrasing implies more certainty than the patient
      expressed

  Inaccurate:
    The event misrepresents what the patient said,
    attributes something they did not say, or gets a key
    detail wrong (e.g., wrong medication name, wrong
    date, opposite sentiment).

RELEVANCE | Would a health coach benefit from remembering
this for future sessions?

  Essential:
    A coach needs this to provide safe, effective care.
    Examples:
    - Current goals, targets, and confidence levels
    - Health conditions, symptoms, or physical limitations
      affecting activity
    - Medications (current, stopped, or changed)
    - Major barriers to progress

  Useful:
    Helpful context that enriches coaching but isn't
    critical. Examples:
    - Preferred walking times or locations
    - Work schedule or lifestyle context
    - One-time events (travel, family events) that
```

```
      affected a specific week
    - Patient preferences or communication style

  Not Useful:
    Not worth storing as a persistent memory. Examples:
    - Trivial pleasantries or acknowledgments
    - Information already obvious from context
    - Overly specific details unlikely to matter later
      (e.g., exact step count on one day)

NOTES (per event)

Use the Notes column on each memory event row to add any
clarification | e.g., why you rated it Partially
Accurate, or additional context the event should have
included.

MISSING MEMORIES & SESSION NOTES

After each session's events, you'll see two prompt rows:
  - "Anything important from Session X NOT captured
    above?" | type your response in the merged cell to
    the right of the prompt.
  - "General notes on Session X?" | same thing, type in
    the merged cell to the right.

Focus on things a coach would want to remember | don't
worry about trivial details.

IMPORTANT NOTES
  - Sessions labeled "Clinical" are simulated clinical
    check-ins (not real conversations). Please still
    evaluate the memory events from these sessions.
  - Read sessions in order | context from earlier
    sessions matters.
  - There are no right or wrong answers. We want your
    professional judgment.
```

Appendix G. Memory Extraction Prompt

The following prompt is provided as system instructions to the extraction LLM (GPT-4o) along with the current memory state formatted with inline IDs. The LLM returns a structured `MemoryDelta` object containing explicit insert, update, and delete operations.

```
Extract important information from the conversation to
remember for future sessions.

=== CORE PRINCIPLE: STABLE FACTS vs. EVOLVING STATE ===

Every piece of information falls into one of three types.
```

```
The type determines whether it gets its own memory or is
consolidated with others.

**Stable facts** | things about the patient that remain
true regardless of how their goals change. Each stable
fact gets its OWN separate memory. NEVER fold a stable
fact into a larger evolving-state memory, because it will
be lost when that memory updates.
Examples of stable fact categories (DO NOT use these
specific examples as extracted memories | only extract
what the patient actually says):
- Equipment owned (e.g., exercise equipment at home)
- Job or daily routine context (e.g., active job that
  contributes to step count)
- Exercise habits or repertoire beyond the current goal
- Preferences (e.g., communication preferences, coaching
  style preferences)
- Preferred locations for activity
- Coping strategies or backup plans

**Evolving state** | things that change over time. These
belong in a single memory per topic that gets UPDATED as
the state changes. Include brief progression history.
Examples of evolving state:
- Current goal, target days, confidence level, and active
  barriers | all in one memory, updated each time any of
  these change
- When the goal changes, update the existing memory to
  reflect the new goal while noting the progression

**Significant events** | notable things that happened at
a specific point in time. Each gets its own memory with
the date.
Examples of significant events:
- Illness or injury that affected the program
- Travel or time away
- Life events that disrupted routine

IMPORTANT: Only extract information the patient actually
stated in the conversation. NEVER insert information from
these instructions or examples.

**Key test**: If the patient's current goal changes next
week, would this fact STILL be true? If yes -> stable
fact -> separate memory. If no -> evolving state ->
update the existing goal memory.

=== DATES ON EVERYTHING ===

Every memory must include a date indicating when the
information was first reported or last changed. The
```

```
session date is provided | use it.

- Stable facts: include the date first mentioned,
  e.g., "(reported [session date])"
- Evolving state: include the date of the latest change,
  e.g., "as of [session date] (changed from X)"
- Events: include when it happened,
  e.g., "[description] on/around [session date]"
- NEVER use relative time ("last week", "yesterday")
  without an absolute date.

=== WHEN TO UPDATE vs. INSERT vs. DELETE vs. NO CHANGE ===

- Evolving state changed -> add an UPDATE with the
  existing memory's memory_id
- New stable fact or new event -> add an INSERT
- Two existing memories overlap on the same evolving
  state -> UPDATE one, DELETE the other
- Nothing new -> return empty lists for all three
  operations

UPDATE and DELETE operations MUST reference a valid
memory_id from the existing memories list. Do NOT
fabricate IDs. If you want to modify a memory but cannot
find its ID, use INSERT instead.

=== DUPLICATE PREVENTION ===

Before adding an INSERT, carefully scan the EXISTING
MEMORIES list above. If an existing memory already
captures the same fact, preference, or event | even if
worded slightly differently | do NOT insert a duplicate.
Instead:
- If the existing memory is accurate and complete ->
  do nothing (return empty lists)
- If the existing memory needs updating -> use UPDATE
  with its memory_id

Never insert a new memory for evolving state that an
existing memory already tracks. But DO insert separate
memories for new stable facts, even if they are
thematically related to an existing evolving-state memory.

=== EVOLVING STATE GRANULARITY ===

For evolving state, group closely-coupled fields into one
memory. Keep unrelated state separate.

- GOOD: One memory for "current step goal + target days +
  confidence + active barriers" | these change together
  and form a coherent program snapshot.
```

```
- BAD: Separate memories for "step goal is 5,000" and
  "goal days are Mon/Wed/Thu" and "confidence is 10/10"
  | these are tightly coupled state about the same
  program.
- BAD: One memory combining "step goal state" and
  "dietary goal state" | these are independent evolving
  states.

Summarize trends rather than listing individual data
points: "consistently exceeds goal, typically 6K-9K
daily steps" NOT each day's count.

=== FAITHFULNESS ===

- Preserve the user's original meaning. Record what they
  said, not a reinterpretation.
- Aspirations and intentions must stay distinguishable
  from current behavior. "I want to do X" != "does X".
- Only extract information explicitly stated | do not
  infer.

=== CATEGORIES ===

Assign one category per memory (for organization only):
- **preference**: Name/nickname, communication style,
  language preference
- **health**: Symptoms, conditions, physical limitations
- **lifestyle**: Diet, exercise, work schedule, sleep
  patterns, habits
- **medication**: OTC medications, supplements
- **fact**: Other important facts

=== NO CHANGES ===

If the conversation contains NO new information worth
remembering, return empty lists for inserts, updates,
and deletes.
```

Appendix H. Reconciliation Engine Prompt

The following prompt is provided to the reconciliation LLM (GPT-4o) along with the patient's extracted memory and the curated clinical summary. The memory content, category, and full clinical summary are inserted into the template at runtime.

```
You are a clinical reconciliation engine. Given a
patient's memory from a health coaching conversation and
their clinical record summary, determine whether there is
a clinically relevant discrepancy.

## Patient Memory
```

```
Content: "{content}"
Category: {category}

## Clinical Record Summary
{clinical_summary}

## Classification Types

- **contradiction**: The memory directly conflicts with
  the clinical record. Example: patient denies a
  documented allergy, claims they stopped a medication
  that is still active, or denies a diagnosed condition.

- **gap_patient**: The memory contains a specific
  clinical assertion that is absent from the clinical
  record AND is clinically actionable. This includes:
  undocumented medications or supplements the patient is
  taking, unreported symptoms, self-reported condition
  changes (e.g., "my anemia resolved years ago" when
  FHIR still shows active), or a specific activity plan
  that poses safety risk given their documented
  conditions (e.g., planning HIIT with active cardiac
  disease).

- **agreement**: The memory is consistent with the
  clinical record. The patient's statement aligns with
  or confirms what is documented.

- **no_fhir**: The memory has no specific clinical
  relevance. This is the correct classification for
  routine coaching content: step goals, walking targets,
  exercise schedules, daily routines, confidence levels,
  preferences, communication style, and general lifestyle
  facts. These should be classified as no_fhir even if
  the patient has conditions broadly related to exercise
  or diet (e.g., a step goal is no_fhir even if the
  patient has prediabetes with an exercise therapy care
  plan | the patient is following their plan, not
  conflicting with it). A memory is only clinically
  relevant if it makes a specific assertion about
  medications, symptoms, diagnoses, allergies, a concrete
  activity that is contraindicated by their conditions,
  or a specific barrier to following a named clinical
  recommendation (e.g., "can't follow the DASH diet due
  to work schedule" references a specific care plan and
  is gap_patient, but "too busy to walk today" is
  no_fhir).

## Instructions
1. First, determine if the memory has specific clinical
   relevance. If it describes routine coaching activity
```

```
   (goals, steps, preferences, schedules, barriers),
   classify as no_fhir and stop | do not search for
   tenuous connections to care plans or conditions.
2. If clinically relevant, compare against the clinical
   summary.
3. Consider data freshness | older clinical records may
   be stale.
4. Assess clinical severity: low, medium, high.
5. For each relevant clinical resource, provide its
   resource_type, code_system, code_value, and display
   exactly as shown in the clinical summary brackets
   [SYSTEM:CODE].
6. Provide a confidence score (0.0-1.0) and brief
   justification.
```

Appendix I. LLM-as-Judge Prompts

I.1. GT Event Matching

For each ground-truth memory event, an LLM judge (GPT-4o) evaluates whether the information is captured in the system's memory state at the corresponding point in the conversation:

```
You are evaluating whether a memory extraction system
captured a specific piece of information from a coaching
conversation.

## Expected Information (Ground Truth)
The patient said: "{utterance}"
This should have been remembered as: "{memory_content}"

## Memory State After Processing That Utterance
These are ALL memories the system had at that point in
time:

{memories}

## Task
Determine if the expected information is captured in the
memory state. The information may be spread across
multiple memories | that is fine. Look across ALL
memories in combination, not just a single one.

- MATCH: The expected information is fully captured
  across the memory state (may use different words, may
  be split across multiple memories)
- PARTIAL: Only some of the expected information appears
  in the memories (key details are missing entirely, not
  just rephrased)
- NO_MATCH: None of the expected information appears in
  any memory
```

```
Respond with valid JSON only, no markdown:
{
  "verdict": "MATCH|PARTIAL|NO_MATCH",
  "content_fidelity": null or 1-5
    (5=perfect capture, 1=barely related),
  "reasoning": "brief explanation"
}
```

I.2. Transcript Judge

A separate judge (GPT-4.1) evaluates the final memory state holistically against the complete conversation transcript across all sessions:

```
You are evaluating a memory extraction system for a
health coaching application. You will read the full
conversation transcript between a coach and patient,
then evaluate the final memory state the system produced.

## Conversation Transcript
{transcript}

## Final Memory State
{memories}

## Task
Score the memory extraction on these 2 dimensions
(1-5 each):

1. **Faithfulness** (1-5): Are the memories faithful to
   what was actually said? No hallucinated, fabricated,
   or distorted information?
   - 5: All memories accurately reflect the conversation,
     no hallucinations
   - 3: Mostly accurate, some minor distortions or
     unsupported inferences
   - 1: Significant inaccuracies or hallucinated
     information

2. **Deduplication** (1-5): Is there one memory per
   concept? No redundant or overlapping memories?
   - 5: No duplicates, each memory is unique
   - 3: Some overlap but mostly distinct
   - 1: Heavy duplication, same info repeated

Respond with valid JSON only, no markdown:
{
  "faithfulness": { "score": 1-5, "reasoning": "..." },
  "deduplication": { "score": 1-5, "reasoning": "..." },
  "overall_notes": "any additional observations"
}
```

Appendix J. Clinical Adjudication Details

A domain expert with a clinical background independently reviewed a stratified random sample of 30 reconciliation scenarios (12.3% of 243), balanced across severity levels (10 low, 10 medium, 10 high) and safety-criticality (15 yes, 15 no). The annotator was provided with each scenario's conversational context, patient statement, and curated clinical summary, and assigned severity (low/medium/high) and safety-critical (yes/no) labels without access to the LLM-generated labels.

For safety-criticality, all six disagreements were cases where the clinician rated a scenario as safety-critical that the LLM did not; the LLM never overestimated safety criticality. For severity, exact agreement was 40.0% (12/30). The clinician rated no scenario as low severity, upgrading all LLM-rated low scenarios to medium or high. This systematic underestimation is consistent with the engine treating care plan non-adherence scenarios as minor informational gaps, whereas clinical judgment recognizes these as meaningful barriers warranting attention.

LLM Usage Disclosure

Large language models were used to assist with code development and manuscript editing during the preparation of this work. All LLM-generated content was reviewed and verified by the authors.