EVOMEDAGENT: AUDITABLE REASONING AND FAILURE-DRIVEN MEMORY EVOLUTION FOR MEDICATION RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medication decisions often involve multiple conditions, ongoing therapies, and patient-specific safety constraints. Effective medication recommendations therefore must go beyond generating clinically plausible outputs to support auditable and safe real-world use. This calls for two complementary capabilities that, however, are often overlooked in existing methods: (1) auditability, enabling clinicians to trace how patient evidence and competing treatment needs lead to individual decisions; and (2) correctability, enabling the system to identify recurrent recommendation errors in comparable clinical contexts, derive targeted corrections, and validate them before updating for safe decisions. We introduce EvoMedAgent, an agentic medication recommender that couples an auditable reasoning pipeline with failure-driven memory evolution. The pipeline makes medication selection auditable through evidence-grounded unit decisions, patient-level reconciliation, and independent review, linking these judgments to each final medication in a reasoning trace. Using this trace and recorded medication sets, the evolution process attributes observed errors, proposes clinically scoped corrections, and validates them before incorporating them into memory. Extensive experiments show that EvoMedAgent achieves substantial improvements over state-of-the-art baselines on real-world datasets, with ablations supporting both mechanisms, and human and LLM evaluations supporting improved auditability. Our code is available at `https://anonymous.4open.science/r/EvoMedAgent/`.

1 INTRODUCTION

Medication recommendation supports clinicians in selecting safe and effective treatments for individual patients (Ali et al., 2023; Elhaddad & Hamam, 2024). Such decisions often involve multiple conditions, ongoing therapies, and patient-specific safety constraints (Wu et al., 2025b; Zhao et al., 2025a). An inappropriate recommendation may leave a treatment need unmet or expose a patient to avoidable harm (Chen et al., 2023). These stakes make it important for clinicians to understand how a recommendation was reached before relying on it, and for errors identified in prior cases to inform decisions for patients with comparable needs (Cooper et al., 2025).

Medication recommendation has evolved from coarse set prediction toward fine-grained, reasoning-aware systems. Earlier deep methods typically predicted medication sets from diagnoses, procedures, and visit histories, often in coarse ATC spaces (Ali et al., 2023; Zhen et al., 2026; Li et al., 2026b). Subsequent longitudinal, graph-, molecular-, and knowledge-enhanced models incorporated treatment histories, drug interactions, and biomedical structure (Zhao et al., 2024b; He et al., 2025; Zhang et al., 2025a; Li et al., 2025; Wu et al., 2025a). These methods improved prediction, but the evidence behind medication choices and the handling of competing needs generally remain implicit in their latents. More recently, LLM-based approaches have targeted fine-grained medication prediction. FLAME improves list-wise alignment through supervised and reinforcement learning (Fan et al., 2025), RES-MR introduces risk-aware reasoning (Wang et al., 2026a), and SafeRx-Agent coordinates knowledge-grounded agents for safe and explainable recommendation (Wang et al., 2026b). Other medical agents coordinate clinical evidence and treatment needs (Xu et al., 2026; Chen et al., 2026; Ge et al., 2026). A broader discussion appears in Appendix A. Nevertheless, rationales or safety checks alone do not show how patient evidence and competing treatment needs

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

shaped medication selection as decisions were made. These methods also rarely explain why recommendation errors recur in comparable clinical contexts or how to propose targeted corrections.

First, clinical auditability helps clinicians judge whether a recommendation accounts for the patient's treatment needs and safety constraints before using it (Zavaleta-Monestel et al., 2026; Alu & Oluwadare, 2026). Here, auditability means being able to trace a recommended medication through the patient evidence and intermediate judgments that shaped its selection, including how competing considerations and subsequent revisions were handled. Consider a patient with a recent coronary stent and active gastrointestinal bleeding. If the final regimen includes an antiplatelet drug, an accompanying rationale may cite the stent, but it does not by itself establish whether the bleeding concern was considered or how it affected the choice. An auditable reasoning process makes these judgments part of medication selection and exposes how competing considerations were reconciled or revised, allowing clinicians to audit the decision path (Cosenza et al., 2026).

Second, observed recommendation failures can guide clinically targeted corrections when their causes and the involved patients are understood. Across patients sharing a comorbidity, repeated omissions of the same treatment may reflect an overlooked treatment objective, a relevant medication never considered, or a safety constraint applied too broadly. These causes require different corrections. If the cause is misidentified, the omission may persist or a correction may affect patients for whom it is inappropriate. Training on historical prescriptions can improve prediction, but parameter updates alone cannot determine which clinical failure was corrected or where a correction should apply (Fan et al., 2025; Wang et al., 2026a). Knowledge-grounded agent workflows can explain individual decisions, yet do not carry corrections from observed failures into later cases (Wang et al., 2026b). Learning from failures therefore requires attributing observed errors, identifying recurrent patterns, and proposing and validating clinically targeted corrections before reuse.

To address these needs, we propose EvoMedAgent, an agentic medication recommender that couples an auditable reasoning pipeline with failure-driven memory evolution. The pipeline organizes each visit into clinical problem units, grounds local medication judgments in patient evidence, reconciles competing needs, and subjects the patient-level decision to independent review and any necessary correction. Its reasoning trace links each recommended medication to the evidence and decisions that support it. Memory evolution uses the trace and observed medication set to attribute observed failures, propose clinically scoped memory updates, and validate their utility in applicable cases before reuse. Together, these processes make recommendations auditable and use failure feedback to improve later decisions. In summary, our contributions are fourfold:

- We propose **EvoMedAgent**, to our knowledge the first agentic medication recommender to couple an auditable reasoning pipeline with validated, failure-driven memory evolution.
- We design an auditable reasoning pipeline to let clinicians examine how patient evidence and competing treatment needs shape each medication choice. It integrates evidence-grounded decisions, reconciliation, and independent review into medication selection.
- We develop failure-driven memory evolution to address recurrent recommendation errors in comparable clinical contexts. It attributes failures and validates clinically scoped memory corrections before they guide later decisions.
- Extensive experiments on MIMIC-III/IV show that EvoMedAgent improves Jaccard by 36.20%/33.24% and F1 by 28.08%/24.99% over the best baselines. Ablations support both mechanisms, while human and LLM evaluations show substantially improved auditability.

2 Problem Formulation

For patient i, $\mathbf{a}_i$ denotes demographic attributes and $V_i^{1:t-1} = \left[\mathbf{v}_i^1, \ldots, \mathbf{v}_i^{t-1}\right]$ represents the visits preceding a target visit t. Each historical visit is represented as $\mathbf{v}_i^\tau = (\mathcal{D}_i^\tau, \mathcal{P}_i^\tau, \mathcal{M}_i^\tau)$, comprising diagnoses, procedures, and medications. The target input is $X_i^t = (\mathbf{a}_i, \mathcal{D}_i^t, \mathcal{P}_i^t, \mathcal{L}_i^t, \mathcal{N}_i^t)$, where $\mathcal{L}_i^t$ contains current laboratory observations and $\mathcal{N}_i^t$ the pre-treatment context summaries. The standardized drug-name vocabulary is denoted by $\mathcal{M}$, and $Y_i^t \subseteq \mathcal{M}$ is the medication set recorded for the target visit. The task is to infer Y_i^t from the target input and longitudinal EHR history (Li et al., 2026b; Fan et al., 2025):

$$\widehat{Y}_i^t = F\big(X_i^t, V_i^{1:t-1}\big) \subseteq \mathcal{M}, \tag{1}$$

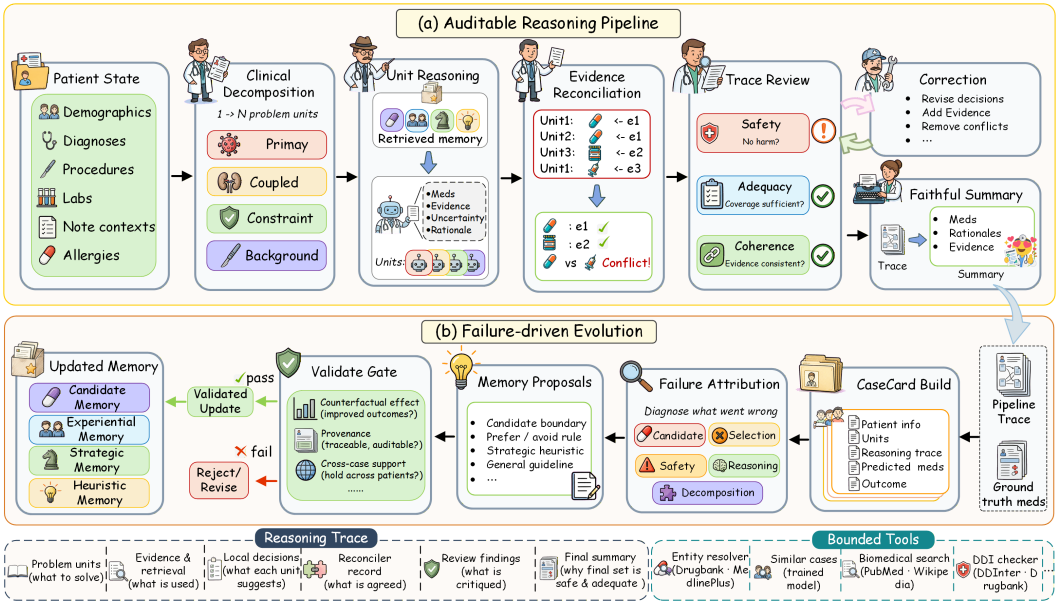

Figure 1: Overview of EvoMedAgent. (a) The auditable reasoning pipeline grounds unit-level medication judgments in patient evidence, reconciles them across clinical problems, and independently reviews the patient-level decision. The reasoning trace links each final medication to its supporting judgments and preserves the reviewed set in the summary. (b) Failure-driven memory evolution attributes observed recommendation errors, groups related failures, and proposes scoped memory updates. Validated updates are incorporated into memory for reuse in comparable cases.

where F can be instantiated by a predictive model or an agentic system. To avoid target leakage, $\mathcal{N}_i^t$ contains five pre-treatment summaries: *Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Family History*, and *Allergies*. The experimental information boundary is in Appendix C.2.

3 EvoMedAgent

EvoMedAgent integrates two complementary processes, an *auditable reasoning pipeline* and *failure-driven memory evolution*, as illustrated in Figure 1. The auditable reasoning pipeline maps a label-free patient state to a medication set through clinical decomposition, memory-grounded unit reasoning, patient-level reconciliation, and independent review. Failure-driven memory evolution analyzes reasoning traces after feedback becomes available, attributes observed recommendation errors, proposes targeted memory updates, and validates their empirical utility before updating the memory state. Together, these processes enable EvoMedAgent to produce evidence-traceable medication recommendations and refine its reasoning knowledge through validated feedback. Appendix B summarizes the notation and overall procedure.

3.1 Auditable Reasoning Pipeline

The auditable reasoning pipeline makes evidence-grounded local judgments, cross-unit reconciliation, and independent review part of medication selection rather than a post hoc account of the final set. To make this process controllable and its responsibility boundaries explicit, EvoMedAgent implements it as a coordinated workflow of specialized LLM agents and deterministic operators. Each LLM agent has a stage-specific responsibility, communicates through schema-constrained records, and may invoke only role-authorized tools within a fixed interaction budget.

EvoMedAgent realizes the pipeline through five coordinated stages. *(1) Clinical decomposition agent.* It organizes the patient state into clinical problem units and verifies that the relevant diagnoses and procedures are represented. *(2) Unit-reasoning agents.* Operating on individual units, these agents produce local medication judgments together with their supporting evidence. *(3) Reconciliation agent.* It consolidates the unit-level judgments into a patient-level medication decision

and resolves inconsistencies across units. *(4) Review and correction.* Three review agents independently assess the reconciled decision for hard safety, clinical adequacy, and evidence coherence. When they identify an actionable deficiency, a constrained correction agent revises the decision within the available candidate boundary, after which the revised result is reviewed again. *(5) Summary agent.* It presents the final reviewed medication set and its supporting rationale without altering the decision. This organization exposes how local judgments, patient-level integration, and independent review jointly shape the final medication set. Deterministic operators validate entity coverage, memory applicability, evidence provenance, and set consistency before intermediate outputs are passed downstream. We denote this constrained reasoning process by $\mathcal{R}$, which jointly produces a medication set and its structured reasoning trace:

$$\begin{aligned} \left(\widehat{Y}_i^t, \mathcal{T}_i^t\right) &= \mathcal{R}\big(X_i^t, V_i^{1:t-1}, \mathcal{K}\big), \\ \mathcal{T}_i^t &= \left(\mathcal{U}_i^t, \mathcal{E}_i^t, \mathcal{J}_i^t, \mathcal{Z}_i^t, \mathcal{H}_i^t, \mathcal{S}_i^t\right). \end{aligned} \tag{2}$$

Here, $\mathcal{K}$ is the memory state available at inference. The trace $\mathcal{T}_i^t$ records clinical problem units $\mathcal{U}_i^t$, retrieved evidence $\mathcal{E}_i^t$, unit-level judgments $\mathcal{J}_i^t$, reconciled decisions $\mathcal{Z}_i^t$, review findings $\mathcal{H}_i^t$, and the final summary $\mathcal{S}_i^t$. Its typed links define a provenance relation $\overset{\mathcal{T}}{\rightsquigarrow}$ from source evidence to the final medication set. EvoMedAgent enforces the following trace invariants:

$$\begin{aligned} \text{(I1) Evidence traceability:} \quad & \forall m \in \widehat{Y}_i^t,\ \exists e \in \mathcal{E}_i^t : e \overset{\mathcal{T}_i^t}{\rightsquigarrow} m, \\ \text{(I2) Summary fidelity:} \quad & \mathrm{Med}\big(\mathcal{S}_i^t\big) = \widehat{Y}_i^t. \end{aligned} \tag{3}$$

The first invariant requires each final medication to retain a path from patient evidence through the judgments and review outcomes that led to its selection. The second requires the summary to faithfully report the final medication set without adding or removing medications.

Clinical Decomposition and Unit Reasoning. Given the label-free patient state, the decomposition agent organizes the current diagnoses and procedures into clinical problem units. A clinical problem unit is a local decision context centered on an anchor diagnosis or procedure. It groups related diagnoses, procedures, and bounded contextual evidence, including current laboratory observations, that should be considered jointly for a local medication decision. These units define local decision boundaries rather than mutually exclusive disease partitions, since clinically coupled units may share supporting entities and are reconciled at the patient level. Each unit contains an anchor, related clinical entities, bounded contextual evidence, and a role in $\{$*primary*, *comorbidity-coupled*, *constraint*, *background*$\}$. The decomposition agent does not make medication decisions. For each clinical problem unit, entity-preserving retrieval identifies candidate medications, applicable experiential rules, and strategic guidance while retaining the identity and role of their supporting entities. A unit-reasoning agent then makes medication-level selection or exclusion judgments within the retrieved candidate boundary and records the supporting evidence, uncertainty, and local rationale. The clinical problem units are processed independently and may be executed in parallel.

Patient-Level Evidence Reconciliation. Local decisions cannot be merged by union because duplicated evidence can create artificial consensus and a medication omitted locally may remain well supported at the case level. EvoMedAgent aggregates every option with its supporting units, distinct clinical entities, history, rules, procedure roles, and negative conflicts, counting identical evidence only once. An optional pre-reconciliation calibration tool provides keep/add suggestions over candidate medications but does not modify the set. The reconciliation agent jointly reviews local selections, calibrated suggestions, and a bounded set of candidate recovery options. It also revises the affected unit decisions so that each recommendation remains aligned with its evidence and rationale. A deterministic set merger subsequently performs only union, deduplication, and conflict recording.

Review, Correction, and Faithful Summarization. After reconciliation, three review agents examine the patient-level decision from complementary perspectives: *(1) Hard safety.* This review checks allergies, absolute contraindications, and patient-specific high-risk interactions, interpreting interaction evidence in the clinical and temporal context of the visit. *(2) Clinical adequacy.* This review checks whether the principal treatment objectives are covered, whether important problems

are untreated or treated unnecessarily, and whether decisions remain consistent across clinical problem units. *(3) Evidence coherence.* This review checks whether each medication, rationale, uncertainty statement, and supporting evidence form a mutually consistent decision path. Their structured findings are passed to a constrained correction agent, which addresses actionable deficiencies only within the available candidate boundary. Any revised decision is submitted to the same review gates again, whereas a decision without an actionable correction proceeds unchanged. Finally, the summary agent explains the latest reviewed medication set, its supporting evidence, and unresolved risks without adding or removing medications. These reviews document potential risks rather than establish clinical safety or replace clinician judgment. Further details are provided in Appendix D.

3.2 Failure-Driven Memory Evolution

Failure-driven memory evolution uses observed errors to develop corrections for comparable future cases while preventing unsupported changes from entering reusable memory. A completed reasoning trace records not only the predicted medication set but also the clinical problem units, evidence, intermediate decisions, and review findings needed to explain a later error. EvoMedAgent uses this feedback through three linked operations. *(1) Structured long-term memory* defines scoped stores that can be retrieved for the relevant clinical problem or reasoning module. *(2) Failure attribution and memory proposal* maps observed discrepancies to explicit failure types and proposes localized updates. *(3) Update validation* evaluates each proposal before it can influence subsequent inference. This organization separates memory scope, proposal rationale, and update evidence.

The three operations are applied to a batch of completed cases rather than to an isolated prediction. At evolution round r, let $\mathcal{I}_{\mathrm{fb}}^{(r)}$ index the patient–visit pairs for which both the reasoning trace and the observed medication set are available. Their batch-level evolution is summarized as:

$$\begin{aligned}
\mathcal{C}_{\mathrm{fail}}^{(r)} &= \mathrm{Attribute}\Big(\big\{(\mathcal{T}_i^t, Y_i^t)\big\}_{(i,t)\in\mathcal{I}_{\mathrm{fb}}^{(r)}}\Big), \\
\mathcal{Q}^{(r)} &= \mathrm{Propose}\Big(\mathcal{K}^{(r)}, \mathcal{C}_{\mathrm{fail}}^{(r)}\Big), \\
\mathcal{A}^{(r)} &= \Big\{q \in \mathcal{Q}^{(r)} : \mathrm{Validate}(q, \mathcal{C}_{\mathrm{learn}}) = 1\Big\}, \\
\mathcal{K}^{(r+1)} &= \mathrm{Update}\Big(\mathcal{K}^{(r)}, \mathcal{A}^{(r)}\Big).
\end{aligned} \tag{4}$$

Here, $\mathcal{C}_{\mathrm{fail}}^{(r)}$ denotes the typed failure records obtained from the completed trace–outcome pairs. $\mathcal{Q}^{(r)}$ contains scoped memory proposals, $\mathcal{C}_{\mathrm{learn}}$ denotes the complete training corpus, and $\mathcal{A}^{(r)}$ contains only proposals that satisfy the corresponding criteria. The four lines represent failure attribution, proposal generation, update validation, and memory update, respectively.

Structured Long-Term Memory. To prevent accumulated experience from diluting the reasoning context or causing context collapse, EvoMedAgent organizes long-term memory into functionally distinct and clinically scoped stores that support selective retrieval. This organization reflects the heterogeneous roles of medication knowledge, separating the definition of viable candidates, condition-specific prescribing experience, reasoning strategies, and module-level guidance. *Candidate memory* specifies the exact medications that may be considered for a diagnosis or procedure, without treating prevalence as a hard cutoff. *Experiential memory* stores condition-matched preferences and avoidance relations with explicit entity ownership. *Strategic memory* provides ordered reasoning guidance for eligible anchors instead of prescribing a fixed drug list. *General heuristic memory* stores sparse module-level guidance for decomposition, unit reasoning, review, or correction. Each entry is retrieved only when its clinical entity or module scope matches the current reasoning context. Bootstrap initialization uses training cases to construct entity-owned candidate associations and conservative experiential rules. These associations define the candidate boundary, not a diagnosis–drug ground-truth mapping. Experiential rules are screened for empirical support, provenance, set-level counterfactual utility, and safety before reuse.

Failure Attribution and Memory Proposal. After the reasoning pipeline is complete and the observed medication set becomes available, EvoMedAgent consolidates the prediction, trace, clinical problem units, evidence, review findings, and medication-set discrepancies into a CaseCard. This

case-level record is passed to the *Failure Attribution Agent*, which may produce multiple FailureCards, each describing one clinically scoped discrepancy and assigning it to exactly one of five types: *(1) Candidate failure.* The relevant candidate boundary does not contain a medication required by the observed outcome. *(2) Selection failure.* An available candidate is omitted or an inappropriate medication is included. *(3) Safety failure.* The decision leaves an actionable medication-level or pairwise safety issue unresolved. *(4) Reasoning failure.* The evidence, rationale, uncertainty, or patient-level reconciliation does not support the decision. *(5) Decomposition failure.* The clinical problem units do not represent an entity or treatment objective needed for the decision. FailureCards with related clinical targets and failure types are grouped into evolution units, allowing recurrent patterns to inform scoped proposals when multiple cases share the same failure. Candidate failures target the candidate boundary, selection and safety failures inform experiential memory, reasoning failures inform strategic memory, and decomposition failures inform module-level heuristics. After each evolution unit is routed to its target memory layer and entity- or module-specific scope, unambiguous updates are instantiated through predefined construction rules, while the *Memory Proposal Agent* handles updates that require contextual judgment. The resulting proposals are then passed to the corresponding validation procedure described next.

Update Validation. Update validation prevents case-specific or weakly supported corrections from propagating into reusable memory. All proposals undergo provenance, scope, and mutation-consistency checks, followed by layer-specific validation. Candidate updates require valid clinical ownership and medication vocabulary. Strategic and general updates must remain consistent with their source cases and intended reasoning scopes. Experiential proposals additionally require empirical support and positive set-level counterfactual utility on training cases matching their clinical scope. For proposal q, let $n_s(q)$ denote its supporting cases, $\mathrm{conf}(q)$ its empirical confidence, $\Delta J(q)$ and $\Delta F_1(q)$ its performance changes, and $\mathrm{Prov}(q)$ its provenance validity. A regular experiential proposal is validated only when:

$$\begin{aligned}\mathrm{Validate}_{\mathrm{exp}}(q) = 1 \iff\ & n_s(q) \geq n_{\min} \ \wedge\ \mathrm{conf}(q) \geq \tau \ \wedge\ \Delta J(q) > 0 \\ & \wedge\ \Delta F_1(q) \geq 0 \ \wedge\ \mathrm{Prov}(q) = 1.\end{aligned} \tag{5}$$

This low-cost set-level counterfactual evaluation estimates direct effects on cached predictions, and validated updates affect only subsequent full-pipeline executions. A failed but revisable proposal is returned to the *Memory Proposal Agent* for bounded revision and reevaluation. Only proposals that pass all applicable checks are validated and retained in memory. All others are rejected. Further details are provided in Appendix E.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on MIMIC-III v1.4 and MIMIC-IV v2.2, two widely used public EHR resources for clinical decision-support research (Johnson et al., 2016; 2020). MIMIC-III contains de-identified critical-care records collected at Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). MIMIC-IV is its successor and provides a larger, more recent, and modular collection of clinical and ICU records from the same institution (Johnson et al., 2020; 2023). Following the fine-grained medication recommendation benchmark established in previous work (Li et al., 2026b; Zhen et al., 2026), we preserve standardized drug-name labels rather than grouping medications into coarse ATC-3 or ATC-4 categories. Both datasets use patient-level training, validation, and test splits of 4:1:1. Cohort statistics appear in Appendix C.1.

Baselines. We compare EvoMedAgent with four groups of baselines, each controlling for a distinct source of improvement. Medication recommenders establish task-specific predictive references (Choi et al., 2016; Shang et al., 2019; Yang et al., 2021; 2023; Li et al., 2025; Wu et al., 2025a; Fan et al., 2025). LLM controls isolate foundation-model knowledge and prompted reasoning (Wei et al., 2022). Medical agents test clinical orchestration and evolving memory (Chen et al., 2026; Wang et al., 2026b; Liang et al., 2026; Shetty et al., 2026; Ge et al., 2026; Sun et al., 2026), whereas recommendation agents assess the transfer of domain-independent collaboration and tool use (Wang et al., 2024a; Yang et al., 2026). All adapted methods follow the same patient-level split,

Table 1: Fine-grained drug-name recommendation on the MIMIC-III and MIMIC-IV test splits. All values are percentages. Higher Jaccard/F1 and lower DDI are better. Bold and underline denote the best and second-best results. Parenthesized values report EvoMedAgent's relative gains over the strongest baseline for Jaccard and F1.

Category	Method	MIMIC-III			MIMIC-IV		
		Jaccard↑	F1↑	DDI↓	Jaccard↑	F1↑	DDI↓
Medication Recommenders	RETAIN (2016)	27.97	42.48	7.35	29.96	45.04	6.49
	GAMENet (2019)	21.43	34.44	7.37	20.99	33.91	6.39
	SafeDrug (2021)	25.64	39.81	6.91	26.16	40.59	6.12
	MoleRec (2023)	28.77	43.63	6.56	27.99	42.89	5.77
	MR-DTR (2025)	28.97	43.43	7.16	34.99	50.67	6.43
	ARMR (2025a)	29.69	44.24	7.26	33.62	48.98	6.58
	FLAME (2025)	34.07	49.34	7.17	35.15	51.85	6.60
LLM Controls	Vanilla LLM	21.54	33.45	10.23	22.91	35.33	9.14
	CoT (2022)	24.61	39.00	9.06	27.16	41.77	8.27
Medical Agents	RareAgents (2026)	30.43	45.13	<u>5.67</u>	32.68	47.69	<u>5.34</u>
	SafeRx-Agent (2026b)	31.13	46.11	**5.63**	32.51	47.11	**5.21**
	D2MDT (2026)	29.00	43.57	7.71	31.51	46.28	6.80
	MARC v1 (2026)	31.49	44.75	8.20	30.84	43.19	7.06
	ClinicalAgents (2026)	27.86	41.31	7.47	28.67	41.76	6.68
	SkeMex (2026)	<u>35.69</u>	<u>50.11</u>	6.48	35.41	49.14	5.89
Recommendation Agents	MACRec (2024a)	21.53	32.98	9.10	18.75	29.58	9.04
	PRTA (2026)	34.58	49.52	8.02	<u>37.66</u>	<u>52.42</u>	5.48
Proposed Method	**EvoMedAgent**	**48.61** *(+36.20%)*	**64.18** *(+28.08%)*	7.18	**50.18** *(+33.24%)*	**65.52** *(+24.99%)*	6.41

prediction-time information boundary, drug-name vocabulary, and evaluation protocol. Individual methods and their task adaptations are detailed in Appendix C.3.

Evaluation Metrics. Following recent medication recommendation studies (He et al., 2025; Fan et al., 2025; Wu et al., 2025a), we report patient-macro Jaccard and F1 to evaluate agreement between the predicted and observed medication sets, together with DDI rate as a medication-interaction proxy. Jaccard measures set overlap relative to the union, while F1 balances patient-level precision and recall. DDI rate measures the proportion of interacting pairs within each recommended set and should not be interpreted as confirmed concurrent exposure. PRAUC is not reported because EvoMedAgent returns a discrete medication set rather than full-vocabulary prediction scores. Formal definitions of the reported metrics are provided in Appendix C.4.

Implementation Details. The patient state comprises demographics, current diagnoses and procedures, laboratory observations, five pre-treatment summaries, and prior visits. The input policy is fixed before evaluation and excludes target medications, post-treatment and discharge documentation, and features whose timing may depend on medication exposure. Bootstrap memory and evolution use training patients only. The learned auxiliary tools are fitted on training patients, configured on validation patients, and fixed before test inference. All LLM calls use `gpt-5.6-terra` with medium reasoning effort. The Trace Review–Correction and memory-proposal revision budgets are both one round. Further implementation details are provided in Appendix H.

4.2 Main Results

Table 1 reports the fine-grained medication-name results on both held-out test sets. EvoMedAgent achieves the highest Jaccard and F1 on MIMIC-III and MIMIC-IV. Relative to the strongest baseline on each dataset, it improves Jaccard and F1 by 36.20% and 28.08% over SkeMex (Sun et al., 2026) on MIMIC-III, and by 33.24% and 24.99% over PRTA (Yang et al., 2026) on MIMIC-IV. The substantial margins on both cohorts demonstrate that the improvement is not confined to a particular dataset. Test-time repetition results are reported in Appendix C.5.

The non-agent results show a progression from fixed set prediction to task-aligned LLM generation. Conventional recommenders map longitudinal EHRs to latent patient representations and fixed set decoders, offering limited support for case-specific decomposition and evidence reconciliation. FLAME's fine-grained list-wise alignment (Fan et al., 2025) outperforms all such models on both datasets. Yet Vanilla LLM and the CoT control (Wei et al., 2022) remain weaker, showing that

Table 2: Component ablations on the MIMIC-III test split. All values are percentages.

Category	Variant	Jaccard↑	F1↑	DDI↓
Full Model	**Full EvoMedAgent**	**48.61**	**64.18**	**7.18**
Auditable Pipeline	Holistic reasoning	41.92	57.62	7.75
	w/o patient-level reconciliation	44.36	60.05	8.16
	w/o Trace Review–Correction	48.40	63.97	7.42
Memory Evolution	Pipeline only (no long-term memory)	39.76	55.02	7.59
	Static memory (no evolution)	43.73	59.36	8.65
	w/o typed failure attribution	45.78	61.42	7.43
	Unified memory (w/o scoped routing)	45.16	60.88	7.55
	w/o corpus-level update validation	44.49	60.14	7.88

language modeling benefits require task-specific alignment rather than model scale or prompted reasoning alone. The stronger agent results point to further gains from structured clinical orchestration.

The agent results show that this potential depends on how reasoning and memory are organized. RareAgents (Chen et al., 2026) and D2MDT (Liang et al., 2026) emphasize multidisciplinary deliberation, MARC v1 (Shetty et al., 2026) preserves intermediate context through staged coordination, and ClinicalAgents (Ge et al., 2026) combines search with dual memory. Their uneven performance, generally above direct LLM controls but not consistently above FLAME, indicates that generic clinical orchestration alone is insufficient. SafeRx-Agent (Wang et al., 2026b) adds medication-specific grounding and safety verification but does not learn from recurrent errors. SkeMex (Sun et al., 2026) evolves utility-governed hierarchical skills from interaction trajectories, whereas EvoMedAgent attributes medication-set errors to typed failure sources, routes recurrent patterns to scoped memories, and validates updates before reuse. The resulting margin is consistent with the benefit of failure-specific evolution over generic skill accumulation. Among recommendation agents, MACRec (Wang et al., 2024a) obtains limited gains from general collaboration, while PRTA (Yang et al., 2026) becomes the strongest MIMIC-IV baseline by invoking trained medication recommenders. EvoMedAgent surpasses both, consistent with the value of medication-specific agentic reasoning and failure-driven memory evolution.

Finally, DDI serves as a supporting result. EvoMedAgent remains comparable to established medication recommenders, although it does not achieve the lowest rates. Because each MIMIC visit-level set aggregates therapies across the visit, we interpret DDI as a conservative risk proxy rather than confirmed concurrent exposure. Appendix C.4 provides a detailed discussion and outlines time-resolved safety evaluation.

4.3 Component and Evolution Analyses

Table 2 isolates the two principal mechanisms of EvoMedAgent. Because evaluating most variants requires rerunning the reasoning pipeline, memory evolution, or both, we conduct the component analysis on MIMIC-III to manage the associated API cost. Holistic reasoning replaces decomposition, unit reasoning, and reconciliation with one patient-level decision. The second variant retains clinical problem units and local decisions but aggregates them without reconciliation. The third stops after reconciliation. Each variant also removes tools specific to its omitted stages. Memory variants disable persistent memory, freeze the bootstrap memory, or remove typed attribution, scoped routing, and corpus-level validation. All comparisons use matched conditions.

The pipeline ablations consistently support structured clinical reasoning. Holistic reasoning yields the lowest recommendation accuracy within this group, showing that a single patient-level decision cannot adequately organize heterogeneous treatment objectives. Removing reconciliation also degrades both recommendation quality and interaction safety because independently reasonable unit decisions may duplicate evidence or conflict at the patient level. Trace Review–Correction minimally affects set accuracy but improves DDI, consistent with its conservative role as a safety gate.

The memory results distinguish knowledge storage from learning through observed failures. Static memory improves recommendation accuracy over the memory-free pipeline, confirming the value of reusable clinical experience, but remains inferior to the evolved memory. Typed attribution and scoped routing are necessary to convert errors into relevant rather than diffuse updates, while corpus-level validation prevents unsupported experience from propagating to later cases. Together, these ablations show that structured reasoning and validated memory evolution improve recommendation quality. Additional evolution, long-tail, and case-level analyses appear in Appendix F.

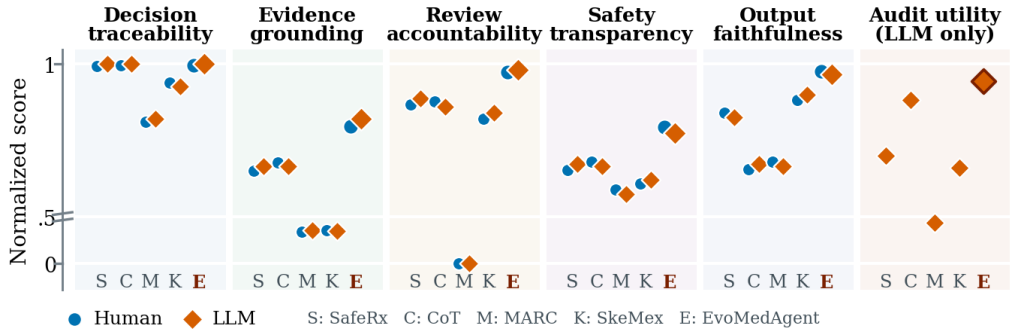

Figure 2: Matched physician- and LLM-based auditability evaluation on the same 50 blinded MIMIC-III cases. Circles and diamonds denote the mean physician and LLM scores, respectively. Audit utility is evaluated by the LLM judge only. All scores are normalized to $[0, 1]$.

4.4 Human- and LLM-Based Auditability Evaluation

Recommendation accuracy does not establish whether a decision can be independently reconstructed. We therefore assess five semantic properties of the recorded reasoning: *traceability* follows each final decision through its responsible stages and material changes, *evidence grounding* connects decisions to addressable patient facts and their provenance, *review accountability* records how critiques affect decisions, *safety transparency* preserves risks and unresolved concerns, and *output faithfulness* checks consistency between the trace and final report. A complementary *audit utility* score measures whether the trace supports concrete localization and tamper-detection tasks.

We evaluate CoT (Wei et al., 2022), MARC v1 (Shetty et al., 2026), SafeRx-Agent (Wang et al., 2026b), SkeMex (Sun et al., 2026), and EvoMedAgent on the same 50 MIMIC-III test cases. Each native trace is mapped without imputation into a common dossier and presented under a blinded system identifier. The LLM judge and two physicians independently score the same dossiers using a shared four-level rubric, with the human result averaged across both physicians (Zhang et al., 2025b; Croxford et al., 2025). Audit utility is evaluated against programmatically specified answers. Further protocol and scoring details appear in Appendix G.

Human and LLM assessments exhibit closely aligned comparative patterns, with only minor score differences. EvoMedAgent attains the strongest overall profile, tying the best LLM traceability and leading the remaining semantic dimensions as well as audit utility. Its larger margins in evidence, review, safety, and faithfulness are consistent with explicit evidence ownership, recorded review dispositions, constrained corrections, and set-preserving summarization. In contrast, the near-saturated traceability results show that a visible decision path alone does not guarantee auditable provenance or terminal consistency. CoT remains strong on operational utility despite weaker output faithfulness, while SafeRx-Agent and SkeMex preserve substantial review structure but expose less patient-specific provenance. MARC v1 records no auditable revision event, which explains its weakness on revision accountability rather than recommendation quality itself.

5 Conclusion and Limitations

We presented EvoMedAgent, an agentic medication recommender that couples auditable clinical reasoning with failure-driven memory evolution. Its pipeline links each final medication to evidence-grounded unit judgments, reconciliation, and independent review, while its evolution mechanism develops clinically scoped updates from attributed errors and validates them before reuse. Experiments on MIMIC-III and MIMIC-IV demonstrate consistent gains over neural, LLM-based, and agentic baselines, with component analyses and human–LLM evaluation supporting its improvements in recommendation quality and auditability.

Despite these results, the evidence remains retrospective and institution-specific. Safety is measured mainly through unordered DDI, while auditability and memory evolution are evaluated on 50 cases and one complete evolution run, respectively. Future work will evaluate EvoMedAgent in prospective multi-institutional settings, with time-resolved safety outcomes and repeated evolution across model configurations. Appendix I provides further discussion.

AI USE STATEMENT

Generative AI tools were used to assist with language editing of the manuscript and with code editing and refactoring. All AI-assisted code changes were reviewed and tested by the authors, and all AI-assisted text was reviewed and revised by the authors. Generative AI was not used to determine the research methodology, experimental results, or scientific conclusions. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses de-identified MIMIC-III and MIMIC-IV data under the applicable PhysioNet access and data-use requirements. The study is retrospective and involves no intervention or interaction with patients. The human auditability evaluation consists solely of expert assessment of blinded retrospective case dossiers and does not involve patient care or the collection of sensitive personal information from the evaluators. EvoMedAgent is evaluated only as a research prototype for clinical decision support and not as an autonomous prescribing system.

REPRODUCIBILITY STATEMENT

Reproducibility details are provided in Section 4.1 and Appendices C–I. Appendix C documents data construction, splits, prediction-time information boundaries, baselines, and recommendation evaluation. Appendices D–E specify the reasoning and memory-evolution procedures. Appendix F provides supplementary analyses, Appendix G specifies the auditability evaluation, and Appendix H reports implementation settings and computational cost. Appendix I discusses the scope and limitations of the present study. The accompanying anonymized implementation provides the prompts, schemas, and code needed to reproduce the pipeline with authorized access to MIMIC data. A companion interactive demonstration visualizes nine de-identified reasoning traces and one failure-to-memory evolution chain at `https://evomedagent-demo.anonymous-page.workers.dev/`.

REFERENCES

Zafar Ali, Yi Huang, Irfan Ullah, Junlan Feng, Chao Deng, Nimbeshaho Thierry, Asad Khan, Asim Ullah Jan, Xiaoli Shen, Wu Rui, et al. Deep learning for medication recommendation: A systematic survey. *Data Intelligence*, 5(2):303–354, 2023.

Fidelis Fidelis Alu and Sunkanmi Oluwadare. An auditable and source-verified framework for clinical AI decision support: Integrating retrieval-augmented generation with data provenance. *Frontiers in Artificial Intelligence*, 9:1737532, 2026.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. RareAgents: Autonomous multidisciplinary team for rare disease diagnosis and treatment. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 101–109, 2026.

Zhao Chen, Ning Liang, Haili Zhang, Huizhen Li, Yijiu Yang, Xingyu Zong, Yaxin Chen, Yanping Wang, and Nannan Shi. Harnessing the power of clinical decision support systems: Challenges and opportunities. *Open Heart*, 10(2):e002432, 2023.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3512–3520, 2016.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.

Jenny Cooper, Shamil Haroon, Francesca Crowe, Krishnarajah Nirantharakumar, Thomas Jackson, Leah Fitzsimmons, Eleanor Hathaway, Sarah Flanagan, Tom Marshall, Louise Jackson, et al. Perspectives of healthcare professionals on the use of artificial intelligence to support clinical decision-making in the management of multiple long-term conditions: Interview study. *Journal of Medical Internet Research*, 27:e71980, 2025.

Bartolomeo Cosenza, Alessandro Concas, Alessandro Virzì, and Gabriele Pannocchia. Offline supervisory fuzzy control for weekly set-point adjustment in home-based congestive heart failure therapy. *Information Sciences*, pp. 123285, 2026.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025.

Ivens da Silva Portugal, Paulo Alencar, and Donald Cowan. Agentic recommender systems: A systematic literature review. *IEEE Transactions on Software Engineering*, 2026.

Malek Elhaddad and Sara Hamam. AI-driven clinical decision support systems: An ongoing pursuit of potential. *Cureus*, 16(4), 2024.

Chenxiao Fan, Chongming Gao, Wentao Shi, Yaxin Gong, Zihao Zhao, and Fuli Feng. Fine-grained list-wise alignment for generative medication recommendation. *arXiv preprint arXiv:2505.20218*, 2025.

Zhuohan Ge, Haoyang Li, Yubo Wang, Nicole Hu, Chen Jason Zhang, and Qing Li. ClinicalAgents: Multi-agent orchestration for clinical decision making with dual memory. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2*, pp. 10969–10980, 2026.

Ming He, Yongjie Zheng, Changle Li, and Man Zhou. Improving the safety of medication recommendation via graph augmented patient similarity network. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 833–842, 2025.

Chaeyoung Huh, Hyunmin Hwang, Jung Hwan Shin, Sungyang Jo, Jinse Park, and Jong Chul Ye. PACE-RAG: Patient-aware contextual and evidence-constrained RAG for clinical drug recommendation. *arXiv preprint arXiv:2603.17356*, 2026.

A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. MIMIC-IV (version 2.2), 2020.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes, 2023. Version 2.2.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. CliCARE: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31554–31562, 2026a.

Yishuo Li, Qi Zhang, Wenpeng Lu, Xueping Peng, Weiyu Zhang, Jiasheng Si, Yongshun Gong, and Liang Hu. Time-aware medication recommendation via intervention of dynamic treatment regimes. In *Proceedings of the ACM on Web Conference 2025*, pp. 5161–5172, 2025.

Yishuo Li, Qi Zhang, Shoujin Wang, Weiyu Zhang, Jiasheng Si, Hu Zhang, and Wenpeng Lu. Medication mapping and diagnosis enhancement for fine-grained medication recommendation. *Information Sciences*, pp. 123930, 2026b.

Yongqi Liang, Qidong Liu, Chunze Yang, Lei Wu, Jiusong Ge, Ni Zhang, and Chen Li. D2MDT: Department-aware multidisciplinary team consultation with deliberation for efficient clinical prediction. *arXiv preprint arXiv:2606.03543*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. MedLA: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 845–853, 2026.

Reza Yousefi Maragheh and Yashar Deldjoo. The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems. *arXiv preprint arXiv:2507.02097*, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Qiyao Peng, Hongtao Liu, Hua Huang, Jian Yang, Qing Yang, and Minglai Shao. A survey on LLM-powered agents for recommender systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 11574–11583, 2025.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. GAMENet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 1126–1133, 2019.

Saisha Shetty, Satvik Tripathi, Austin Lin, Colin Zhao, Theodore Kim, Don Enwerem, Jacinta Arnold, Shahriar Faghani, and Tessa S Cook. MARC v1: An open-source multi-agent framework for clinical AI reasoning and coordination. *arXiv preprint arXiv:2608.13476*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 8634–8652, 2023.

Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258, 2024.

Cong Wang, Jin Li, Shoujin Wang, Yishuo Li, Huilin Gu, and Wenpeng Lu. RES-MR: Risk-aware reasoning for explainable and safe medication recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1834–1845, 2026a.

Xinyu Wang, Hanwei Wu, Zhenghan Tai, Sicheng Lyu, Qincheng Lu, Ziyu Zhao, Jijun Chi, Jingrui Tian, Xiao-Wen Chang, and Ziyang Song. SafeRx-Agent: A knowledge-grounded multi-agent framework for safe and explainable medication recommendation. *arXiv preprint arXiv:2605.29146*, 2026b.

Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. MACRec: A multi-agent collaboration framework for recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2760–2764, 2024a.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 24824–24837, 2022.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

Feiyue Wu, Tianxing Wu, and Shenqi Jing. ARMR: Adaptively responsive network for medication recommendation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pp. 7831–7839, 2025a.

Jialun Wu, Kai He, Rui Mao, Xuequn Shang, and Erik Cambria. Harnessing the potential of multimodal EHR data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion*, 123:103283, 2025b.

Guoli Xiong, Zhijiang Yang, Jiacai Yi, Ningning Wang, Lei Wang, Huimin Zhu, Chengkun Wu, Aiping Lu, Xiang Chen, Shao Liu, et al. DDInter: An online drug–drug interaction database towards improving clinical decision-making and patient safety. *Nucleic Acids Research*, 50(D1): D1200–D1207, 2022.

Gelei Xu, Xueyang Li, Yixiong Chen, Yuying Duan, Shuqing Wu, Haoxinran Yu, Ching-Hao Chiu, Juntong Ni, Ningzhi Tang, Toby Jia-Jun Li, et al. A comprehensive survey of AI agents in healthcare. *Journal of Biomedical Informatics*, pp. 105045, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, pp. 17577–17604, 2025.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. SafeDrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 3735–3741, 8 2021.

Mingdai Yang, Zhiwei Liu, Weizhi Zhang, Yibo Wang, Hao Peng, and Philip Yu. Personalized recommendation tool learning via autonomous language agents. *arXiv preprint arXiv:2607.19739*, 2026.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. MoleRec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pp. 4075–4085, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Esteban Zavaleta-Monestel, José A Castro-Gamboa, and Jeaustin Mora-Jiménez. Artificial intelligence should not prescribe alone: A risk-stratified pharmacy governance framework for agentic clinical systems. *Cureus*, 18(8), 2026.

Haodi Zhang, Jiawei Wen, Jiahong Li, Yuanfeng Song, Liang-Jie Zhang, and Lin Ma. SSPNet: Leveraging robust medication recommendation with history and knowledge. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pp. 9465–9473, 2025a.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4888–4914, 2025b.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *Proceedings of the 14th International Conference on Learning Representations*, pp. 86069–86100, 2026.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1315–1324, 2017.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 19632–19642, 2024a.

Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li. Enhancing precision drug recommendations via in-depth exploration of motif relationships. *IEEE Transactions on Knowledge and Data Engineering*, 2024b.

Chuang Zhao, Hui Tang, Hongke Zhao, and Xiaomeng Li. Beyond sequential patterns: Rethinking healthcare predictions with contextual insights. *ACM Transactions on Information Systems*, 43 (4), 2025a.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. Leave no patient behind: Enhancing medication recommendation for rare disease patients. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–542, 2024c.

Zihao Zhao, Chenxiao Fan, Chongming Gao, Fuli Feng, and Xiangnan He. Addressing overprescribing challenges: Fine-tuning large language models for medication recommendation tasks. *arXiv preprint arXiv:2503.03687*, 2025b.

Yeguang Zhen, Shoujin Wang, Yishuo Li, Liang Hu, Defu Lian, Cong Wang, and Wenpeng Lu. EHR-based medication recommendation: A stage-oriented survey and unified benchmark, 2026. Preprint.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 46595–46623, 2023.

Yinghao Zhu, Lei Gu, Zixiang Wang, Haoran Sang, Dehao Sui, Wen Tang, Lan Mi, Yasha Wang, Junyi Gao, Liang Yao, Tianfan Fu, Ewen Harrison, Lequan Yu, and Liantao Ma. Auditing medical multi-agent AI reveals risks of false consensus. *arXiv preprint arXiv:2510.10185*, 2025.

A Related Work

A.1 EHR-Based Medication Recommendation

EHR-based medication recommendation has progressed by enriching patient-state modeling, medication knowledge, and prescription decision making. Early methods encoded diagnoses and procedures from the current visit (Zhang et al., 2017), while longitudinal models incorporated prior visits to capture disease progression and treatment continuity (Choi et al., 2016; Yang et al., 2021; 2023; Li et al., 2025). More recent methods have increasingly supplemented these formulations with patient similarity, co-prescription patterns, drug–drug interactions, molecular structures, and biomedical knowledge (Zhao et al., 2024b; He et al., 2025; Zhang et al., 2025a; Li et al., 2025; Wu et al., 2025a). LLM-based methods further introduce richer clinical semantics and fine-grained set generation. FineMed (Li et al., 2026b) constructs disease-oriented sub-recommendations through LLM-assisted preprocessing, FLAME (Fan et al., 2025) aligns generated lists with target medication sets, and RES-MR (Wang et al., 2026a) incorporates risk-aware reasoning into medication selection. PACE-RAG (Huh et al., 2026) retrieves cases around patient-specific clinical features and constrains prescription refinement with the resulting treatment patterns. SafeRx-Agent (Wang et al., 2026b) is closest to our setting, coordinating specialty agents with external knowledge and safety-aware revision to produce traceable recommendations. Its workflow, however, does not use observed recommendation failures to revise reusable memory. EvoMedAgent extends this agentic direction by making the end-to-end decision process auditable and coupling it with failure-driven memory evolution.

A.2 LLM Agents for Decision Making

LLM agents extend language models from passive generation to sequential decision making through planning, reasoning, tool use, and memory (Yao et al., 2023; Maragheh & Deldjoo, 2025; Peng et al., 2025; da Silva Portugal et al., 2026). Recent clinical agents emphasize process reliability: MedLA (Ma et al., 2026) supports premise-level correction, CliCARE (Li et al., 2026a) grounds EHR decisions in clinical guidelines, and MedAgentAudit (Zhu et al., 2025) audits evidence loss and false consensus. Their focus is the reliability of a clinical decision rather than learning from observed medication outcomes across patients. In recommendation, MACRec (Wang et al., 2024a) coordinates specialized agents for analysis and reflection, whereas PRTA (Yang et al., 2026) treats conventional recommenders as tools that an LLM planner selects and refines for individual users. A parallel line enables agents to learn without parameter updates. Reflexion (Shinn et al., 2023) stores linguistic feedback from previous trials, ExpeL (Zhao et al., 2024a) abstracts cross-task experience into reusable natural-language knowledge, and more recent methods (Xu et al., 2025; Wang et al., 2024b) organize experience as reusable workflows or dynamically linked memories. These approaches demonstrate the value of non-parametric experience, but their updates are generally optimized for task completion, personalization, or retrieval quality. They do not require each update to be attributed to a medication-specific failure and validated across a broader patient population. Moreover, retaining a reasoning trace does not by itself establish clinical auditability, which requires explicit evidence provenance and consistency throughout the decision process.

B Notation and Algorithmic Overview

Table 3 summarizes the notation used in the task formulation and EvoMedAgent. Algorithm 1 presents the corresponding reasoning and memory-evolution processes. The prediction stage never accesses the target medication set. Observed outcomes enter only the subsequent evolution stage.

C Data, Baselines, and Evaluation Protocol

This section specifies the empirical basis of the comparison. It first describes cohort and label construction, then separates prediction-time evidence from outcome feedback, and finally documents baseline adaptation and recommendation evaluation. This separation ensures that no post-decision information can contribute to a reported recommendation.

Table 3: Key notation used throughout the paper.

Symbol	Description
Task and input	
i, t, τ	Patient index, target-visit index, and historical-visit index.
$\mathbf{a}_i$	Demographic attributes of patient i.
$\mathcal{D}_i^t, \mathcal{P}_i^t, \mathcal{L}_i^t, \mathcal{N}_i^t$	Current diagnoses, procedures, laboratory observations, and pre-treatment context summaries.
$V_i^{1:t-1}$	Longitudinal EHR history preceding target visit t.
X_i^t	Label-free input for the target visit.
$\mathcal{M}$	Vocabulary of standardized medication names.
$Y_i^t, \widehat{Y}_i^t$	Observed and predicted medication sets for the target visit.
Auditable reasoning	
$\mathcal{U}_i^t, \mathcal{E}_i^t, \mathcal{J}_i^t$	Clinical problem units, retrieved evidence, and unit-level judgments.
$\mathcal{Z}_i^t, \mathcal{H}_i^t, \mathcal{S}_i^t$	Reconciled decisions, review findings, and faithful summary.
$\mathcal{T}_i^t$	Structured reasoning trace for patient–visit pair (i, t).
Memory evolution	
$\mathcal{K}^{(r)}$	Long-term memory state at evolution round r.
$\mathcal{I}_{\text{fb}}^{(r)}$	Patient–visit pairs with traces and observed outcomes available at round r.
$\mathcal{C}_{\text{learn}}$	Complete training corpus used to identify eligible cases and evaluate proposals against case-level predictions and observed medication sets.
$\mathcal{C}_{\text{fail}}^{(r)}, \mathcal{Q}^{(r)}, \mathcal{A}^{(r)}$	Typed failures, proposed memory updates, and validated updates at round r.

Table 4: Statistics of the processed fine-grained medication-recommendation cohorts.

Item	MIMIC-III	MIMIC-IV
Patients	6,360	8,949
Visits	16,976	24,106
Train/validation/test patients	4,197/1,081/1,082	5,906/1,521/1,522
Diagnosis codes	4,672	11,036
Procedure codes	1,420	4,810
Standardized drug names	718	874
Avg./max. visits per patient	2.67/42	2.69/30
Avg./max. diagnoses per visit	13.92/39	19.06/39
Avg./max. procedures per visit	3.95/32	3.24/34
Avg./max. medications per visit	20.49/78	24.54/118
Observed DDI rate (%)	8.95	7.04

C.1 Processed Cohorts and Fine-Grained Labels

Following the established fine-grained cohort construction (Zhen et al., 2026), each processed record contains the ordered visits of one patient. The final visit is treated as the target visit, while earlier visits supply longitudinal diagnoses, procedures, and medications. Patients, rather than individual visits, are assigned to non-overlapping training, validation, and test partitions. We preserve standardized drug names as prediction labels and use exact vocabulary matching throughout evaluation. Thus, medications remain distinct even within the same ATC-3 or ATC-4 category.

Table 4 summarizes the resulting cohorts. Diagnosis and procedure vocabulary sizes count standardized codes, whereas the medication vocabulary counts standardized drug names. Consistent with the evaluation metric, the observed DDI rate is computed within each processed visit and then averaged across visits. MIMIC-IV contains more diagnosis codes and a larger medication vocabulary than MIMIC-III, making the fine-grained output space both larger and sparser.

Algorithm 1 EvoMedAgent: auditable inference and failure-driven evolution.

Input: Target input X_i^t, history $V_i^{1:t-1}$, memory $\mathcal{K}^{(r)}$, training corpus $\mathcal{C}_{\text{learn}}$, and optional feedback indexed by $\mathcal{I}_{\text{fb}}^{(r)}$.
Output: Recommendation $\widehat{Y}_i^t$, trace $\mathcal{T}_i^t$, and updated memory $\mathcal{K}^{(r+1)}$.

Stage I: Auditable reasoning
$\mathcal{U}_i^t \leftarrow \text{Decompose}(X_i^t, V_i^{1:t-1})$.
for all $u \in \mathcal{U}_i^t$ **do**
  $\mathcal{E}_{i,u}^t \leftarrow \text{Retrieve}(u, \mathcal{K}^{(r)})$
  $\mathcal{J}_{i,u}^t \leftarrow \text{UnitReason}(u, \mathcal{E}_{i,u}^t)$
end for
$\mathcal{E}_i^t \leftarrow \{\mathcal{E}_{i,u}^t\}_{u \in \mathcal{U}_i^t}$ and $\mathcal{J}_i^t \leftarrow \{\mathcal{J}_{i,u}^t\}_{u \in \mathcal{U}_i^t}$
$\mathcal{Z}_i^t \leftarrow \text{Reconcile}(\mathcal{U}_i^t, \mathcal{E}_i^t, \mathcal{J}_i^t)$.
$h_i^t \leftarrow \text{Review}(\mathcal{Z}_i^t, \mathcal{E}_i^t)$ and $\mathcal{H}_i^t \leftarrow [h_i^t]$
while h_i^t contains an actionable finding and the review budget remains **do**
  $\mathcal{Z}_i^t \leftarrow \text{Correct}(\mathcal{Z}_i^t, h_i^t)$
  $h_i^t \leftarrow \text{Review}(\mathcal{Z}_i^t, \mathcal{E}_i^t)$
  $\mathcal{H}_i^t \leftarrow \text{Append}(\mathcal{H}_i^t, h_i^t)$.
end while
$\widehat{Y}_i^t \leftarrow \text{Med}(\mathcal{Z}_i^t)$ and $\mathcal{S}_i^t \leftarrow \text{Summarize}(\mathcal{Z}_i^t, \mathcal{E}_i^t, \mathcal{H}_i^t)$
$\mathcal{T}_i^t \leftarrow (\mathcal{U}_i^t, \mathcal{E}_i^t, \mathcal{J}_i^t, \mathcal{Z}_i^t, \mathcal{H}_i^t, \mathcal{S}_i^t)$.

Stage II: Failure-driven memory evolution
if $\mathcal{I}_{\text{fb}}^{(r)} \neq \emptyset$ **then**
  $\mathcal{C}_{\text{fail}}^{(r)} \leftarrow \text{Attribute}(\{(\mathcal{T}_i^t, Y_i^t)\}_{(i,t) \in \mathcal{I}_{\text{fb}}^{(r)}})$.
  $\mathcal{Q}^{(r)} \leftarrow \text{Propose}(\mathcal{K}^{(r)}, \mathcal{C}_{\text{fail}}^{(r)})$ and $\mathcal{A}^{(r)} \leftarrow \emptyset$
  for all $q \in \mathcal{Q}^{(r)}$ **do**
    $v \leftarrow \text{Validate}(q, \mathcal{C}_{\text{learn}})$
    while v is revisable and the revision budget remains **do**
      $q \leftarrow \text{Revise}(q, v)$
      $v \leftarrow \text{Validate}(q, \mathcal{C}_{\text{learn}})$
    end while
    if $v =$ VALID **then**
      $\mathcal{A}^{(r)} \leftarrow \mathcal{A}^{(r)} \cup \{q\}$.
    end if
  end for
  $\mathcal{K}^{(r+1)} \leftarrow \text{Update}(\mathcal{K}^{(r)}, \mathcal{A}^{(r)})$.
else
  $\mathcal{K}^{(r+1)} \leftarrow \mathcal{K}^{(r)}$.
end if
return $\widehat{Y}_i^t, \mathcal{T}_i^t, \mathcal{K}^{(r+1)}$.

We retain the established benchmark's patient ordering and contiguous split boundaries.

C.2 Prediction-Time Information Boundary

EvoMedAgent constructs a label-free projection of each target visit rather than serializing the complete processed record. Current diagnoses and procedures retain their exact names and codes, while current laboratory observations are represented as compact name–value–unit strings. Because raw MIMIC notes are lengthy, irregularly structured, and interleave information from different stages of care, directly presenting them to the agents would yield an unstable and potentially leakage-prone context (Johnson et al., 2016; 2023). Following LLM-assisted fine-grained EHR construction (Li et al., 2026b; Zhen et al., 2026), we extract and summarize five pre-treatment components: *Chief Complaint*, *History of Present Illness*, *Past Medical History*, *Family History*, and *Allergies*. Only text under explicitly matched section headings is retained. The summarization prompt further restricts the LLM to this pre-treatment context and explicitly prohibits the inclusion of treatments, medications administered during the visit, or subsequent-care content. Historical visits contribute diagnoses, procedures, and medications, but not clinical notes or laboratory observations.

Table 5: Information boundary across prediction, learning, and evaluation. Prediction-time access applies to every data split.

Information	Permitted use
Pre-treatment patient state and longitudinal history	Prediction-time reasoning in every split and training-side learning.
Post-treatment and discharge text	Excluded from all model inputs.
Training medication labels	Bootstrap memory, failure-driven evolution, and auxiliary-tool fitting, but never prediction-time input.
Validation medication labels	Configuration and decision-threshold selection only, but never prediction-time input or memory evolution.
Test medication labels	Held-out evaluation only.

Current laboratory observations populate $\mathcal{L}_i^t$ and enter the reasoning pipeline as bounded contextual evidence. They may condition a clinical problem unit and its medication decisions, but do not serve as diagnosis or procedure anchors.

Table 5 separates the information available to prediction-time reasoning from labels used for learning or evaluation. The prediction-time boundary applies identically to training, validation, and test cases: target medications never enter the reasoning pipeline that produces their recommendations.

C.3 Baseline Methods

The baseline suite separates four possible sources of improvement: task-specific predictive modeling, foundation-model knowledge, clinical agent organization, and general recommendation-oriented agent design. Published medication recommenders retain their fine-grained benchmark configurations without architectural or task-specific changes.

Representative Medication Recommenders. We retain seven models spanning the principal non-agent design families without listing architecturally similar predictors. **RETAIN** (Choi et al., 2016) is a reverse-time attention model for interpretable longitudinal EHR prediction. **GAMENet** (Shang et al., 2019) augments patient history with EHR co-occurrence and DDI graph memories. **SafeDrug** (Yang et al., 2021) combines molecular graph encoders with a controllable DDI objective, whereas **MoleRec** (Yang et al., 2023) reasons over medication substructures and their patient-conditioned dependencies. **MR-DTR** (Li et al., 2025) represents dynamic treatment evolution with time-aware patient similarity. **ARMR** (Wu et al., 2025a) combines piecewise temporal learning with adaptive response modeling to balance retained and newly introduced medications. **FLAME** (Fan et al., 2025) extends the medication-wise classification formulation of LAMO (Zhao et al., 2025b) into a two-stage generator, using drug-level filtering followed by list-level refinement with supervised fine-tuning and step-wise reinforcement learning.

LLM Controls. **Vanilla LLM** directly predicts the medication set from the same patient state and output vocabulary, without explicit decomposition, persistent memory, or inter-agent communication. **CoT** (Wei et al., 2022) uses the same model and interface but elicits structured step-by-step clinical reasoning before prediction. These controls determine whether improvements can be explained by the medical knowledge of the foundation model or by reasoning prompts alone.

Medical Agents. **RareAgents** (Chen et al., 2026) organizes an autonomous multidisciplinary team with medical tools and dynamic memory, providing a strong control for medical collaboration and memory-enabled treatment reasoning. **SafeRx-Agent** (Wang et al., 2026b) is the closest medication-specific competitor, grounding a multi-agent recommendation process in external knowledge and explicit safety verification. **D2MDT** (Liang et al., 2026) dynamically assigns department experts to an EHR case and uses residual deliberation to resolve outstanding disagreements. **MARC v1** (Shetty et al., 2026) serializes extraction, reasoning, answer generation, and evaluation through role-specialized clinical agents with explicit intermediate context. **ClinicalAgents** (Ge et al., 2026) employs search-based clinical orchestration with working and experience memories to

support hypothesis revision and evidence verification. **SkeMex** (Sun et al., 2026) distills interaction trajectories into governed skill memories without updating model parameters, making it the most direct comparison for non-parametric self-evolution. Together, they test whether generic multidisciplinary coordination, traceable orchestration, or self-evolving medical memory suffices without medication-specific failure attribution and validated updates.

General Recommendation Agents. **MACRec** (Wang et al., 2024a) coordinates a manager, analysts, a searcher, and a reflector for recommendation. We preserve this architecture but replace user–item interfaces with the shared clinical state and medication-set output. **PRTA** (Yang et al., 2026), implemented through RecToolAgent-MR, treats conventional recommenders as selectable tools and uses an LLM planner with reflection to personalize tool use. Together, they test cross-domain transfer without medication-specific provenance, safety review, or failure attribution.

Methods with public implementations follow their released code and reported best configurations, whereas methods without public code are reproduced from the procedures and settings described in the original papers. Each agent baseline retains its original role structure, reasoning protocol, and reported tool support. Only the input and output interfaces are adapted to the shared patient state and standardized medication vocabulary. This protocol evaluates complete agent systems under a common data boundary and evaluation setting without substituting tools that would alter the intended baselines. The resulting comparison is system-level rather than tool-normalized, because replacing specialized tools with a common inventory would change the methods being evaluated.

C.4 Evaluation Metrics and Statistical Analysis

Following standard medication-recommendation evaluation (Shang et al., 2019; Yang et al., 2023; Li et al., 2025), patient-macro Jaccard over the test patient–visit index set $\mathcal{I}_{\text{test}}$ is defined as:

$$\text{Jaccard} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{(i,t)\in\mathcal{I}_{\text{test}}} \frac{|\widehat{Y}_i^t \cap Y_i^t|}{|\widehat{Y}_i^t \cup Y_i^t|}. \tag{6}$$

The per-case value is one only when both sets are empty. Patient-level precision and recall are used to compute macro F1:

$$\begin{aligned} \text{Precision}_i^t &= \frac{|\widehat{Y}_i^t \cap Y_i^t|}{|\widehat{Y}_i^t|}, \\ \text{Recall}_i^t &= \frac{|\widehat{Y}_i^t \cap Y_i^t|}{|Y_i^t|}, \\ \text{F1}_i^t &= \frac{2\,\text{Precision}_i^t\,\text{Recall}_i^t}{\text{Precision}_i^t + \text{Recall}_i^t}, \\ \text{F1} &= \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{(i,t)\in\mathcal{I}_{\text{test}}} \text{F1}_i^t\,. \end{aligned} \tag{7}$$

Precision is zero when no medication is predicted, recall is zero when the reference set is empty, and F1 is zero when their sum is zero.

For a medication set S, $\text{Pairs}(S)$ denotes its unordered medication pairs. Following prior DDI-aware evaluation (Shang et al., 2019; Yang et al., 2021; 2023), the patient-macro DDI rate is computed using the DrugBank-derived binary interaction matrix $\mathbf{A}_{\text{ddi}}$ (Wishart et al., 2018):

$$\text{DDI} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{(i,t)\in\mathcal{I}_{\text{test}}} \frac{1}{|\,\text{Pairs}(\widehat{Y}_i^t)|} \sum_{\{m,m'\}\in\text{Pairs}(\widehat{Y}_i^t)} \mathbf{A}_{\text{ddi}}[m,m']. \tag{8}$$

A prediction with fewer than two medications receives a DDI rate of zero.

Interpreting Visit-Level DDI. The DDI rate computed from an unordered visit-level medication set is not a direct measure of concurrent exposure. Medications recorded within the same visit may be administered at different treatment stages, used as temporary substitutions, or separated by monitoring and washout periods, whereas the set-based metric evaluates every pair as potentially concurrent (Zhao et al., 2025a; Li et al., 2026b). A flagged pair therefore indicates possible interaction

Table 6: Test-time repetition on MIMIC-III for LLM controls and agent-based methods. Values are percentages and show mean ± sample standard deviation across five runs.

Method	Jaccard↑	F1↑	DDI↓
LLM controls			
Vanilla LLM	21.31 ± 0.55	33.10 ± 0.82	10.14 ± 0.45
CoT	24.92 ± 0.51	39.41 ± 0.67	9.06 ± 0.40
Medical agents			
RareAgents	30.51 ± 0.33	45.21 ± 0.41	5.76 ± 0.12
SafeRx-Agent	31.29 ± 0.39	46.32 ± 0.47	5.61 ± 0.12
D2MDT	29.12 ± 0.53	43.71 ± 0.65	7.73 ± 0.20
MARC v1	31.41 ± 0.40	44.67 ± 0.46	8.13 ± 0.19
ClinicalAgents	27.81 ± 0.48	41.24 ± 0.61	7.54 ± 0.35
SkeMex	35.57 ± 0.41	50.02 ± 0.43	6.48 ± 0.11
General recommendation agents			
MACRec	21.62 ± 0.65	33.09 ± 0.81	8.94 ± 0.48
PRTA	34.75 ± 0.34	49.69 ± 0.42	8.04 ± 0.11
EvoMedAgent	$\mathbf{48.67 \pm 0.48}$	$\mathbf{64.24 \pm 0.47}$	7.12 ± 0.17

Table 7: Paired patient-level improvements of EvoMedAgent over representative agent baselines on MIMIC-III. Gains and 95% confidence intervals are reported in percentage points.

Baseline	Jaccard		F1	
	Δ [95% CI]	p	Δ [95% CI]	p
SafeRx-Agent	17.49 [16.69, 18.30]	< 0.001	18.07 [17.22, 18.92]	< 0.001
RareAgents	18.19 [17.37, 18.98]	< 0.001	19.06 [18.22, 19.91]	< 0.001
D2MDT	19.61 [18.78, 20.43]	< 0.001	20.61 [19.73, 21.47]	< 0.001
SkeMex	12.92 [12.14, 13.72]	< 0.001	14.07 [13.23, 14.89]	< 0.001

risk rather than a confirmed unsafe co-prescription. Conversely, an exceptionally low DDI rate may reflect the omission of clinically necessary therapies rather than a safer or more complete regimen. This limitation is particularly relevant to reasoning-based systems because a trace may distinguish sequential use even though the final unordered set and its DDI score discard that temporal justification. We therefore interpret DDI jointly with Jaccard and F1. Accordingly, DDI non-regression is applied only to safety-oriented proposals and is not treated as a complete clinical-safety objective. A stronger safety assessment would require administration start and end times, dose, frequency, route, treatment duration, interaction severity, and clinician review, which we leave to future time-resolved evaluation.

C.5 Test-Time Repetition and Cohort-Level Uncertainty

Reconstructing the learned state for every repetition is prohibitively expensive for agentic systems, particularly when memory evolves from accumulated experience. We therefore train or construct each system once, freeze its parameters and memory, and repeat MIMIC-III test inference five times under an identical configuration. Every repetition evaluates the identical set of held-out cases. Table 6 summarizes the resulting variation, which reflects stochastic test-time generation rather than variability from rebuilding the learned agent state.

The repetitions above quantify stochastic test-time variation with a fixed learned state. We separately assess uncertainty across the test population by pairing methods on the same patients. Table 7 reports mean gains from EvoMedAgent with 95% confidence intervals obtained from 10,000 paired bootstrap resamples. Two-sided paired t-tests are applied to the corresponding patient-level metric differences.

The confidence intervals remain well above zero for all four comparisons, including SkeMex, the strongest MIMIC-III baseline. These cohort-level results quantify uncertainty across patients and do not estimate variability from independently reconstructing the learned memory state.

Table 8: Role-specific contracts in EvoMedAgent.

Role	Receives	Produces	Authority boundary
Decision construction			
Clinical Decomposition	Current entities, laboratory context, bounded history, and scoped heuristics	Clinical problem units with functional roles	No medication selection or tool use
Unit Reasoning	One clinical problem unit, entity-owned memory, and bounded evidence	Local medication decisions with uncertainty and provenance	No target labels or drugs outside the candidate boundary
Evidence Reconciliation	Local decisions and bounded predictive advice	Coherent patient-level decisions with unit ownership	Predictive advice cannot serve as clinical evidence
Review and communication			
Trace Review	Reconciled decisions, supporting evidence, and any correction proposal	Typed critiques and review verdicts	No direct modification of the medication set
Correction	Actionable critiques and admissible actions	Revised decisions with repaired rationales	No additions or replacements outside the supplied boundary
Faithful Summarization	Final reviewed state and supporting evidence	Fixed-set summary and drug-level rationales	No addition, deletion, or renaming of medications

D Auditable Reasoning Pipeline: Implementation Details

This appendix details how the auditable reasoning pipeline is constrained at prediction time. It first specifies role authority and core prompts, then describes clinical problem unit construction and memory retrieval. The remaining subsections present role-constrained tool access and the validation contracts that preserve trace consistency through summarization.

D.1 Role Contracts and Core Prompts

An auditable workflow requires each reasoning role to have an explicit responsibility and authority boundary. This subsection specifies these role contracts and the instructions that implement them. Table 8 summarizes what each role receives, produces, and may alter. Table 9 then presents the corresponding core prompt templates.

Executable prompts combine role instructions with runtime case fields, structured-output schemas, tool specifications, and repair clauses. Because reproducing them in full would be prohibitively long, Table 9 presents their role-defining instruction templates, while the complete executable prompts are available in the accompanying code repository. Braced terms denote case-specific fields inserted at runtime. Every role must use only supplied information, preserve its assigned authority, and leave material decisions recoverable from the trace.

D.2 Clinical Problem Units and Memory Retrieval

Clinical problem units connect the current patient state to focused medication reasoning. This subsection describes how these units are constructed and validated, how relevant memory is retrieved for each unit, and which information the resulting local decisions must preserve.

Clinical Decomposition must cover every current diagnosis and procedure at least once. It cannot introduce an entity that is absent from the current visit, and it cannot use historical entities as current anchors. Procedure-centered units are reserved for procedures that create a distinct medication objective, such as anesthesia, prophylaxis, hemostasis, ventilation-related sedation, dialysis, or therapeutic transfusion. Routine access and purely diagnostic procedures remain related evidence unless they independently change the medication decision.

For each validated unit, retrieval constructs an entity-preserving memory bundle. Candidate medications, experiential rules, strategic guidance, and general heuristics retain the diagnosis or procedure that supplied them and that entity's role in the unit. Applicability is evaluated before prompt construction using current entities, unit role, demographics, allergy text, and medications from prior visits. Thus, inapplicable rules are withheld from the LLM.

Table 9: Core prompt templates for the auditable reasoning pipeline.

Role	Core instruction
Clinical Decomposition	*Given {current diagnoses and procedures}, {laboratory context}, {pre-treatment context}, and {history}, construct clinical problem units that each define one coherent medication-decision objective. Cover every current diagnosis and procedure using exact supplied names, assign a functional role to each unit, and keep historical facts as context. Do not recommend medications or call tools.*
Unit Reasoning	*For {clinical problem unit}, assess {entity-owned candidates and memory} against {patient evidence} and optional {tool evidence}. Select or avoid only medications within the supplied boundary. For each judgment, identify its evidence, uncertainty, and unit-level rationale. Do not infer target labels or transfer a decision from another unit without unit-specific support.*
Evidence Reconciliation	*Reconcile {unit decisions} at the patient level. Resolve duplicated support, competing ownership, cross-unit conflicts, treatment continuity, and advisory calibration signals. Return a coherent set of final unit decisions in which every retained medication has a responsible unit and supporting evidence. Treat predictive scores as advice rather than clinical evidence.*
Trace Review	*Review {reconciled decisions and trace} from the assigned perspective: hard safety, clinical adequacy, or evidence coherence. Identify only case-specific, evidence-supported deficiencies. For each finding, state the affected decision, severity, evidence, uncertainty, and whether a set change is justified. Do not modify the medication set.*
Correction	*Given {review findings} and the {candidate boundary}, apply only corrections supported by the trace. Preserve unaffected decisions, update the responsible unit and rationale for every accepted change, and retain unresolved concerns. Do not introduce an unsupported medication or silently discard a reviewed finding.*
Faithful Summarization	*Explain {latest reviewed state} without reassessing it. Echo the final medication set exactly, provide one evidence-linked rationale per medication, and retain unresolved safety or uncertainty statements. Do not add, remove, or rename medications.*

Memory retrieval first identifies entries that match the current clinical problem unit and then limits how many are placed in the LLM context. Bootstrap candidate associations are ordered by training support within each entity, while prevalence tiers and matched experiential rules identify higher-priority options. Among the remaining candidates, fallback retrieval prioritizes support from multiple matched entities, then anchor and procedure support, and finally within-entity order. It admits at most six additional medications per matched entity and twelve per clinical problem unit. These limits control context length without deleting stored entries. Table 23 consolidates them with the other inference budgets.

After the memory context has been assembled, each Unit Reasoning Agent returns selected and avoided medications, a discrete certainty level, explicit uncertainty reasons, concise evidence items, and a local rationale. Every selected medication must use an exact standardized name from the candidate boundary, while a background unit may return no medication. The original local output remains available as the provenance record when Evidence Reconciliation later changes the unit's final decision.

D.3 Role-Constrained Tool Access

Tools provide targeted evidence or learned advice that is not contained in the bounded patient state, but they do not make medication decisions. This subsection defines the function of each tool, the roles authorized to invoke it, and how its output is distinguished from patient evidence. The reported tool suite is one implementation of this bounded interface rather than a fixed requirement of the framework. Alternative sources may be substituted provided that their provenance and role-specific authority remain explicit. Table 10 summarizes these constraints.

Access is enforced through stage-specific allowlists. Clinical Decomposition, Faithful Summarization, Failure Attribution, and Memory Proposal cannot invoke the tools in Table 10. The Hard-Safety Review always receives the DDI result, whereas other authorized calls are optional and subject to the interaction budgets in Table 23. Wikipedia is restricted to entity clarification and is not used as evidence for treatment decisions.

Learned auxiliary tools expose bounded suggestions only to their designated stages and never act as standalone recommenders. For training queries, the query patient is excluded from similar-case retrieval, so every returned precedent is drawn from a different training patient.

Table 10: Evidence and learned auxiliary tools used by EvoMedAgent, together with their authorized roles.

Tool	Source and construction	Decision-support function	Authorized roles
External evidence tools			
DDI checker	Local pair index constructed from DDInter and normalized to DrugBank identities (Xiong et al., 2022; Wishart et al., 2018).	Identifies known interactions in a proposed medication set without treating a match as proof of concurrent exposure.	Unit Reasoning when strategy-triggered, Hard-Safety Review as required evidence, and Correction
Entity resolver	DrugBank drug records and MedlinePlus Connect disease records (Wishart et al., 2018).	Resolves unfamiliar entities and aliases without expanding the formal medication vocabulary.	Bootstrap Attribution, Unit Reasoning, Evidence Reconciliation, all three review roles, and Correction
Similar-case retriever	Training-only ARMR medication-response representations (Wu et al., 2025a) combined with note-summary embeddings.	Returns up to three precedents as weak cohort evidence rather than patient facts or mandatory recommendations.	Bootstrap Attribution, Unit Reasoning, Clinical-Adequacy Review, and Correction
Biomedical retriever	Focused PubMed evidence queries and Wikipedia entity lookups returned as concise source records.	PubMed supplies targeted biomedical evidence, whereas Wikipedia is used only to clarify unfamiliar entities and terminology.	Bootstrap Attribution, Unit Reasoning, all three review roles, and Correction
Learned auxiliary tools			
Pre-reconciliation calibration tool	Candidate-level linear and gradient-boosted estimators fitted on training data and selected on validation data (Pedregosa et al., 2011).	Supplies bounded keep/add advice over formal candidates without serving as clinical evidence.	Evidence Reconciliation
Post-reconciliation residual tool	Gradient-boosted residual estimator fitted on training data and selected by validation Jaccard and DDI.	Proposes a candidate-constrained correction for independent review.	Trace Review and Correction

The external knowledge resources are fixed consistently across experiments. DrugBank version 5.1.22 (Wishart et al., 2018) provides drug identities and pharmacological descriptions, while DDInter 2.0 (Xiong et al., 2022) supplies interaction evidence. PubMed, MedlinePlus Connect, and Wikipedia are accessed through their public interfaces.[1] Online results were cached at first retrieval between May and August 2026 and then reused across runs. Wikipedia remains restricted to terminology clarification, and none of these resources exposes cohort outcomes or target medication labels.

D.4 Reasoning Validation Contracts

Reasoning validation preserves the consistency and traceability of the medication decision as it passes between pipeline stages. It does not reassess clinical correctness. Instead, it verifies that every intermediate artifact is well formed, every post-reconciliation medication-set change is supported by an explicit review action, and the final summary faithfully reflects the reviewed state.

These checks operate on structured records. Reconciled medications retain their responsible units, supporting evidence, conflicts, and advisory signals, while review findings identify the affected decision, supporting evidence, severity, and proposed resolution. Free-text comments cannot directly change the medication set. An accepted correction must update the affected ownership, rationale, uncertainty, and evidence before rereview. Table 11 summarizes the checks at each transition. The review budget permits one correction and a second review, including bounded repair of revisable structural errors before the corrected state is reviewed again.

These contracts give the provenance relation in Section 3 an executable interpretation. For every final medication, the stored trace connects its unit-level judgment and patient evidence to the reconciliation decision, any conflicts and corrections, and the latest review outcome. Original local decisions remain available for comparison but are never confused with the current reconciled state.

[1] Official interfaces: PubMed, MedlinePlus Connect, and the Wikipedia API.

Table 11: Deterministic contracts across prediction-time reasoning stages.

Transition	Enforced contract
Clinical Decomposition	Unit identifiers and roles are valid, anchors use exact current entities, and every current diagnosis and procedure is covered.
Retrieval → Unit Reasoning	Retrieved entries retain valid ownership and satisfy applicability conditions. Selected medications remain within the supplied vocabulary and candidate boundary.
Reconciliation → reconciled state	Every reconciled medication is assigned to a unit, duplicated evidence is collapsed, conflicts are recorded, and the reconciled patient set equals the unit union.
Review → Correction	Critiques reference existing decisions and evidence. Proposed actions use available candidates and satisfy severity, provenance, and set-change constraints.
Correction → re-review	The revised unit union reconstructs the proposed set, affected rationales are repaired, and every changed state receives another review.
Reviewed state → summary	The summary preserves the exact set, supplies one drug-level rationale per medication, and cites only evidence present in the trace.

Table 12: Hierarchical organization and internal indexing of long-term memory.

Layer	Stored content	Internal index and filtering
Candidate	Entity-linked medication options	Entity type → exact diagnosis or procedure, with prevalence guiding ranking rather than membership
Experiential	Conditioned medication-set transformations	Owner entity → rules applicable to the patient context and unit role
Strategic	Ordered checks and reasoning guidance	Anchor entity → highest-ranked eligible strategies
General heuristic	Sparse guidance for decomposition, unit reasoning, review, or correction	Reasoning module → optional specialization or entity scope

E Failure-Driven Memory Evolution: Implementation Details

This appendix follows the evolution process from memory organization to failure grouping, scoped proposal formation, layer-specific validation, and memory updating. Each subsection corresponds to one successive stage of this process.

E.1 Memory Organization and Bootstrap Initialization

Unrestricted natural-language reflection becomes increasingly difficult to retrieve as experience accumulates and can introduce irrelevant context (Zhang et al., 2026). EvoMedAgent therefore organizes long-term memory as a coarse-to-fine index. The four functional layers form the top-level partition, while entries within each layer are indexed by exact clinical entities or reasoning scopes. This hierarchy determines where experience is stored and updated. Appendix D describes prediction-time retrieval and context assembly. Table 12 summarizes the memory organization.

Candidate memory records the medications associated with each clinical entity. Every attributed medication is inserted into the corresponding entity-level candidate set, but its presence neither mandates selection nor restricts the medication to a unique entity. Separately, the stored association statistics quantify the strength of the association between medication m and entity e using the denominator-smoothed estimate:

$$p_{\text{smooth}}(m \mid e) = \frac{\text{support}(m, e)}{\text{denominator}(e) + \lambda}, \tag{9}$$

where $\text{support}(m, e)$ counts training cases that associate m with e, $\text{denominator}(e)$ counts training cases containing e, and λ is a conservative smoothing coefficient. We use $\lambda = 2$ to prevent associations supported by very few cases from receiving disproportionate weight. The resulting score controls retrieval priority rather than admission to candidate memory. We designate values of at least 0.40 as strong and values in $[0.30, 0.40)$ as moderate. The remaining candidates stay in memory and can enter the reasoning context through bounded fallback retrieval.

Bootstrap initialization addresses the cold-start problem using training labels. A label-attribution call assigns each observed medication to one or more exact current diagnoses or procedures, with multi-entity ownership permitted for coupled objectives. These associations populate candidate

Table 13: Set-level medication actions used by experiential proposals.

Action	Source regimen	Target regimen	Effect
prefer	empty	nonempty	add absent target medications
avoid	nonempty	empty	remove present source medications
replace	nonempty	nonempty	remove source-only and add target-only drugs

memory. Initial positive experiential rules require repeated support and are capped per entity to prevent prevalence from becoming a ground-truth lookup. Bootstrap artifacts initialize memory but are never presented as patient-specific evidence at prediction time.

On MIMIC-III, the bootstrap candidate memory covers 96.32% of the observed test medication labels on average. After memory evolution, coverage reaches 96.58%, while the average number of candidates per represented entity remains 10.70. This small change shows that bootstrap initialization already establishes a high-coverage candidate boundary. For an entity absent from candidate memory, terminology retrieval may clarify its clinical meaning but does not create an entity–medication association. Candidate medications may still be supplied by other matched entities in the same clinical problem unit, and all recommendations remain within $\mathcal{M}$.

E.2 Failure Attribution and Scoped Proposal Formation

Failure attribution converts observed recommendation errors into localized evidence for memory evolution. Once an observed medication set becomes available, a CaseCard augments the completed reasoning trace with set discrepancies and applied corrections. Missing and extra medications are recomputed from the predicted and observed sets, and all referenced units, entities, candidates, critiques, and actions are checked against the trace. The Failure Attribution Agent may then derive multiple FailureCards from one CaseCard. Cards sharing a failure type, anchor, and clinical target form an evolution unit, revealing recurrent patterns when present and anchoring proposed corrections to their clinical context.

After each evolution unit is routed to its target memory layer and entity- or module-specific scope, proposal formation proceeds in two ways. When the attributed evidence admits a unique update under predefined construction rules, EvoMedAgent instantiates the proposal deterministically. This path covers label-supported candidate additions, entity-coverage corrections, and recurrent context-consistent medication omissions. All remaining evolution units are assessed by the *Memory Proposal Agent*, which assigns one of three first-pass decisions: *propose* a scoped update, return *no change* when current memory already captures the observed pattern, or *defer* when the available cases do not support a stable update. Every deterministically instantiated or agent-generated proposal preserves its source FailureCards, targets exactly one memory scope, and undergoes the validation required by that scope.

Experiential proposals encode one of three medication-set transformations, summarized in Table 13: *prefer* adds a supported target regimen, *avoid* removes a contraindicated or unsupported source regimen, and *replace* exchanges the source-only medications for the target-only medications. Source and target regimens denote joint sets rather than alternatives. Thus, replacing $[A]$ with $[A, B]$ retains A and adds B.

Table 14 presents the core instructions connecting observed outcomes to memory updates. Outcome labels are available only within this learning process and are never exposed to prediction-time roles.

E.3 Layer-Specific Update Validation

Update validation prevents an unsupported or overly broad proposal from becoming persistent memory. It combines shared provenance and scope constraints with evidence requirements specific to each memory layer. Every proposal must reference valid source FailureCards, target exactly one memory layer and internal scope, and encode an action consistent with known clinical entities and medications. A proposal is validated only after both these shared constraints and its layer-specific criteria are satisfied.

Table 14: Core prompt templates for memory initialization and evolution.

Role	Core instruction
Memory initialization	
Bootstrap attribution	*Given {training patient state} and {observed medication set}, assign every medication to the exact current diagnosis or procedure it most plausibly addresses. Permit multiple owners only when distinct entities provide genuine clinical support. Record confidence and rationale, and do not treat this label-driven association as prediction-time evidence.*
Failure-driven evolution	
Failure Attribution	*Analyze {completed CaseCard}. Cover every missing and extra medication, identify the responsible clinical unit and exact target entity, and classify each discrepancy as a candidate, selection, safety, reasoning, or decomposition failure. Emit multiple failures when independently supported, but retain only conditions that explain each error.*
Memory Proposal	*Compare each {evolution unit} with {current scoped memory}. Propose one atomic update, return no change, or defer. Preserve source provenance, use only supplied entities and medications, and use a narrower conditional update when cases imply incompatible behavior. Do not claim empirical support or accept the proposal.*
Memory Proposal (revision)	*Given {proposal} and {validation feedback}, repair only the failed scope or performance condition. Return a revised scoped proposal, no change, or defer. Do not change its target layer, invent supporting evidence, or override a non-revisable rejection.*

For an experiential rule ρ, $n_e(\rho)$ and $n_s(\rho)$ denote its eligible and supporting cases, respectively. Its empirical confidence is:

$$\mathrm{conf}(\rho) = \frac{n_s(\rho)}{n_e(\rho)}. \tag{10}$$

An eligible case contains at least one clinical problem unit that satisfies the rule's entity, role, demographic, and clinical-factor conditions. For efficient validation, the proposal's set transformation is applied to the cached prediction for every eligible case in $\mathcal{C}_{\mathrm{learn}}$. This set-level counterfactual evaluation estimates direct empirical utility without rerunning the downstream reasoning pipeline for each case. $\Delta J(\rho)$ and $\Delta F_1(\rho)$ denote the resulting changes from those predictions, averaged over the same eligible cases.

In our experiments, $\mathcal{C}_{\mathrm{learn}}$ comprises the complete training corpus, with each case retaining the prediction state and observed medication set required for validation. The main-text validation rule uses $n_{\min} = 2$ from two distinct visits and $\tau = 0.50$. Every eligible case contributes to support, confidence, and performance changes, irrespective of whether it also motivated the proposal. This full-corpus criterion allows a recurrent error shared by the matching cohort to be evaluated as a systematic update rather than losing its supporting cases during validation. Related failures are consolidated into evolution units before proposal formation, but each resulting proposal is evaluated separately. Refinements, splits, and merges must not underperform the entries they replace, and semantically overlapping validated rules are consolidated during memory finalization. Candidate prerequisites and source provenance are verified against the stored corpus rather than accepted from the proposal itself. Safety-oriented rules must introduce no detected regression under the monitored DDI criterion. Because eligibility is evaluated over cached predictions, this procedure requires no additional LLM calls. At larger institutional scale, indexed cohort construction can replace exhaustive scanning without changing the validation criterion.

A restricted exception preserves clinically meaningful long-tail evidence. A tentative rule may be validated with one eligible and one supporting case only when it is a *prefer* rule, contains a discriminative condition beyond its owner entity, improves Jaccard without reducing F1, and refers to an existing candidate. *Avoid* and *replace* rules cannot use this exception. A tentative rule therefore represents a narrow empirical pattern rather than a broadly established clinical principle.

The remaining layers use checks matched to their function. Candidate updates require an exact dataset medication and a valid clinical owner. Strategic updates must preserve their anchor scope and remain consistent with their source cases. General heuristics must retain the intended module scope and express behavior that applies across their supporting observations. Incompatible evidence prevents validation rather than being resolved through an unrestricted update.

Proposals remain pending until validation is complete. If a failed criterion is revisable, the validation feedback is returned to the *Memory Proposal Agent*, which may produce a revised proposal for reevaluation within the budget in Table 23. A proposal is validated only after it satisfies all applicable

Table 15: Memory evolution statistics on MIMIC-III. Panel (a) reports first-pass decisions of the Memory Proposal Agent. Panel (b) reports the outcomes of both agent-generated and deterministically instantiated proposals.

(a) First-pass agent decisions

Target layer	**Evolution units**	**Propose**	**Defer**	**No change**
Experiential	29,650	12,847	14,431	2,372
Strategic	1,522	886	635	1
General heuristic	9	9	0	0
Total	31,181	13,742	15,066	2,373

(b) Proposal outcomes

Target layer	**Proposals**	**Revised**	**Validated**	**Rejected**
Candidate	918	0	918	0
Experiential	37,671	1,019	22,443	15,228
Strategic	886	0	761	125
General heuristic	44	0	11	33
Total	39,519	1,019	24,133	15,386

checks. It is rejected when it violates a non-revisable constraint, remains invalid after the revision budget is exhausted, or does not satisfy the support required for validation. By contrast, *no change* and *defer* terminate proposal formation without creating a proposal and therefore do not enter this validation process.

E.4 Evolution Schedule

The evolution schedule determines when observed outcomes may influence memory and prevents them from leaking into the reasoning trace that produced the corresponding recommendation. Candidate associations are initialized from the training partition, after which the reasoning pipeline produces label-free predictions for the same cases. Once outcomes are available, these completed training cases form $\mathcal{C}_{\text{learn}}$, containing 4,197 MIMIC-III cases and 5,906 MIMIC-IV cases. The evolution cycle processes candidate and experiential updates before strategic and general updates, allowing higher-level guidance to depend on validated medication-level experience. The resulting memory state is fixed before validation and test inference. Table 23 reports the batching, context, and revision budgets.

Each recommendation is completed under $\mathcal{K}^{(r)}$ before its outcome can enter the feedback buffer, and validated updates affect only later visits through $\mathcal{K}^{(r+1)}$. The reported evaluation disables test-time evolution so that all test visits use the same memory state.

F Supplementary Experimental Analyses

This section complements the aggregate results with analyses of memory evolution, validation sensitivity, recommendation-set cardinality, long-tail performance, and case-level reasoning.

F.1 Memory Evolution Statistics

Table 15 separates the first-pass decisions of the *Memory Proposal Agent* from the final outcomes of all proposals. Panel (a) records whether an agent-assessed evolution unit produces a proposal, is deferred because its evidence is insufficient, or requires no change to the current memory. Updates with a unique construction under predefined rules are instantiated deterministically and therefore enter Panel (b) without an agent decision. Panel (b) follows all proposals through validation. The *revised* column identifies proposals that undergo bounded revision and is not an additional outcome. A proposal is *validated* only when its final form passes all applicable checks and enters memory, and all remaining proposals are *rejected*.

Table 16: Sensitivity of experiential-rule validation on the MIMIC-III training corpus. Mean confidence is a percentage, and set-level counterfactual gains are percentage points over the eligible cases of each retained rule.

Analysis	$n_{\min}$	τ	Retained rules	Mean conf.	Mean ΔJ	Mean ΔF_1
Confidence threshold	2	0.40	2,478	73.32	1.91	2.86
	2	**0.50**	**2,301**	**75.09**	**1.95**	**2.91**
	2	0.60	1,747	80.48	2.05	3.06
	2	0.70	1,231	86.25	2.14	3.20
Support threshold	**2**	0.50	**2,301**	**75.09**	**1.95**	**2.91**
	3	0.50	2,257	75.00	1.93	2.89
	5	0.50	2,141	74.80	1.89	2.83
	10	0.50	1,798	74.48	1.83	2.74

Table 17: Average number of medications per MIMIC-III test visit. The observed medication set is included as a reference.

Method	Observed	FLAME	SafeRx-Agent	SkeMex	PRTA	EvoMedAgent
Avg. medications	18.51	13.51	14.83	13.17	18.23	18.08

The statistics reveal that experiential memory accounts for most evolution activity. Many experiential units are deferred or terminated without a change, while a substantial fraction of proposals is rejected during validation. By contrast, the smaller strategic and general layers admit comparatively few updates. The resulting pattern is consistent with selective evolution: case-level failure evidence drives most memory updates, whereas higher-level guidance changes less frequently.

F.2 Experiential Validation Sensitivity

We examine whether experiential-memory validation depends strongly on its confidence and support thresholds. Holding the candidate rule pool and training corpus fixed, we reapply the validation criteria while varying τ and $n_{\min}$. Evolved rules must satisfy the same full-corpus support, confidence, and performance criteria under every setting. The rule counts in Table 16 refer to distinct active rules after corpus validation and semantic deduplication, rather than the individual proposals reported in Table 15.

Increasing τ yields a smaller but more consistently supported rule set, together with higher mean set-level counterfactual gains. Increasing $n_{\min}$ reduces retention more gradually while leaving mean confidence comparatively stable. The default setting balances these effects by excluding the lower-confidence rules admitted at $\tau = 0.40$ without imposing the sharper contraction observed under stricter confidence thresholds.

F.3 Recommendation-Set Cardinality

Table 17 compares the average prediction size with the observed medication-set size on MIMIC-III. EvoMedAgent closely matches the observed cardinality, whereas FLAME, SafeRx-Agent, and SkeMex tend to produce smaller sets. PRTA also approaches the observed size, but its lower set overlap in Table 1 shows that cardinality alone does not explain EvoMedAgent's gains.

F.4 Long-Tail Analysis

We apply the same visit-level construction to diagnosis contexts and medication targets using frequencies from the training partition (Zhao et al., 2024c). For an entity set $\mathcal{B}_i^t \in \{\mathcal{D}_i^t, Y_i^t\}$, let $f_{\mathrm{tr}}(z)$ denote the number of training visits containing entity z, and let N_{tr} denote the total number of training visits. Its aggregate rarity is:

$$R_i^t(\mathcal{B}) = \frac{1}{|\mathcal{B}_i^t|} \sum_{z \in \mathcal{B}_i^t} -\log\left(\frac{f_{\mathrm{tr}}(z)+1}{N_{\mathrm{tr}}+1}\right). \tag{11}$$

Table 18: Long-tail performance on MIMIC-III. Diagnosis- and medication-tail cohorts contain 399 and 346 test visits, respectively. Metrics are computed over complete medication sets and reported as percentages.

Method	Diagnosis-tail cohort			Medication-tail cohort		
	Jaccard↑	F1↑	DDI↓	Jaccard↑	F1↑	DDI↓
SafeRx-Agent	31.06	46.04	5.16	31.78	46.92	5.69
RareAgents	29.42	43.89	5.38	30.53	45.34	5.61
D2MDT	27.61	41.93	8.01	25.24	39.27	7.31
SkeMex	34.84	49.27	6.04	36.67	51.25	6.37
EvoMedAgent	**46.87**	**62.60**	7.26	**46.14**	**62.05**	6.94

Table 19: Matched comparison on a representative MIMIC-III test visit. The coverage profile summarizes the method-specific prediction and trace. TP, FP, and FN denote medication counts.

Method	Case-specific coverage profile	TP	FP	FN	Jaccard↑	F1↑
CoT	Concentrates on a small chronic cardiovascular subset and omits most procedure-associated medications.	3	1	20	12.50	22.22
SafeRx-Agent	Recovers additional background therapies but leaves substantial gaps in the operative regimen.	7	5	16	25.00	40.00
SkeMex	Provides the strongest baseline coverage, yet still misses much of the procedure-centered treatment set.	9	3	14	34.62	51.43
EvoMedAgent	Decomposes the operative and longitudinal objectives, reconciles their medications, and retains timing-qualified interaction concerns.	**21**	**2**	**2**	**84.00**	**91.30**

We compute $R_i^t(\mathcal{D})$ from the complete current diagnosis set and $R_i^t(Y)$ from the complete observed medication set. For each basis, the tail threshold is the 75th percentile of its aggregate-rarity scores among training visits. Test visits above this threshold form the *Tail* cohort, and the remainder form the comparison cohort. Performance is evaluated over the complete recommendation set. This construction prevents a single rare entity from determining cohort membership and avoids defining rarity from the test distribution. Table 18 compares EvoMedAgent with representative agent baselines on the two tail cohorts.

EvoMedAgent maintains substantially higher Jaccard and F1 across both tail cohorts. SkeMex is the strongest baseline, especially on medication-tail visits, while SafeRx-Agent and RareAgents remain more stable than D2MDT but achieve lower absolute accuracy. These results support EvoMedAgent's effectiveness on rare clinical contexts and medication profiles without implying uniformly lower DDI.

F.5 End-to-End Case Study

We examine a representative MIMIC-III test visit involving interacting operative and longitudinal treatment objectives. This example illustrates the end-to-end trace rather than providing evidence of aggregate superiority. The patient is a 75-year-old man with recurrent heart failure, severe native and bypass-graft coronary disease, prior myocardial infarctions, an implanted defibrillator, and a need for further cancer surgery. The target visit includes two-vessel coronary bypass and extracorporeal circulation. Pre-treatment context records a high-risk redo-CABG evaluation and perioperative interruption of warfarin, while prior visits provide longitudinal cardiovascular context. Table 19 contrasts the methods on this matched case, and Table 20 shows how EvoMedAgent reaches its result.

For this case, the comparison illustrates the benefit of structured decomposition and reconciliation over flat reasoning or generic medical memory. EvoMedAgent assigns the surgical pathway its own medication objective, reconciles it with longitudinal cardiovascular treatment, and exposes interaction concerns without conflating possible sequential exposure with confirmed coadministration. The resulting trace improves coverage while keeping the remaining omissions, extras, and unresolved safety questions directly inspectable.

Table 20: End-to-end EvoMedAgent trace for the representative MIMIC-III test visit.

Phase	Stage	Trace excerpt
Reasoning	Clinical Decomposition	Forms an active-heart-failure unit, a coupled coronary-disease/CABG unit, constraint units for hypertension and peripheral vascular disease, and background units for arrhythmia and malignancy context.
	Unit Reasoning	Separates chronic cardiovascular continuity from the CABG pathway. Procedure-centered reasoning recovers anesthesia, analgesia, gastrointestinal protection, electrolyte, and operative-preparation medications, while the cardiovascular units retain heart-failure, secondary-prevention, and antiarrhythmic therapies.
	Evidence Reconciliation	Assigns retained drugs to their strongest clinical owners and integrates current procedures, prior visits, entity-owned memory, and calibrated candidate evidence.
Review	Trace Review	Flags possible ketorolac–aspirin, lisinopril–potassium, and amiodarone–simvastatin risks. Because concurrent administration is not established, the trace preserves them as temporal cautions rather than silently deleting the medications.
Output	Faithful Summarization	Returns 23 medications with medication-level rationales. Twenty-one match the observed set. Latanoprost and meperidine are missed, while lisinopril and ketorolac are extra. The summary preserves the review findings and perioperative timing uncertainty.

G Auditability Evaluation Protocol

We randomly sample 50 visits from the MIMIC-III test split and evaluate the same visits for all five methods in Section 4.4. This appendix specifies how their heterogeneous native traces are normalized, blinded, scored, and assessed for inter-rater agreement (Tam et al., 2024; Zhang et al., 2025b). Each trace is mapped to a common representation of medication decisions, supporting evidence, revisions, safety findings, and final outputs. Missing rationales, provenance, reviews, or revisions remain missing rather than being inferred from the recommended medications.

Method names are replaced by fixed random aliases before evaluation. Neither the LLM judge nor the physicians receive target medications, recommendation metrics, or method identities. Additional stages and longer explanations receive no intrinsic credit, so a one-step method can obtain the highest score when its decisions and terminal state are fully reconstructable (Zheng et al., 2023). Table 21 defines this rubric.

The first five dimensions use a shared four-level maturity scale. A score of 0 denotes absent, contradictory, or unauditable information. Scores of 1 and 2 indicate, respectively, weak cues requiring substantial inference and a recoverable path with a material gap in scope, provenance, disposition, or terminal linkage. A score of 3 denotes a complete, case-specific, addressable, and internally consistent record. Every nonzero LLM score must cite a valid trace pointer, and an evidence score of at least 2 must also cite the corresponding prediction-time patient fact. For method m and dimension d, the normalized score is:

$$S_{m,d} = \frac{1}{3N} \sum_{i=1}^{N} s_{i,m,d}, \qquad s_{i,m,d} \in \{0, 1, 2, 3\}. \tag{12}$$

Audit utility is an objective task score rather than a maturity rating. For each dossier, the evaluator must locate one specified medication decision, set change, and safety or uncertainty item when applicable. It must also detect, classify, and exactly localize one programmatically injected inconsistency. Private answers and accepted pointers are generated before evaluation. The case score averages successful applicable tasks, and the reported utility averages the case scores:

$$U_c = \frac{1}{|A_c|} \sum_{q \in A_c} \mathbb{1}[\text{success}_{c,q}], \qquad U = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} U_c. \tag{13}$$

Table 21: Auditability criteria. The LLM judge and physicians assess semantic trace quality, while audit utility is evaluated against predefined answers.

Criterion	Question assessed
Semantic trace quality: LLM and physician evaluation	
Decision traceability	Can a final medication decision be linked to its case scope, responsible stage, material transitions, and terminal status?
Evidence grounding and provenance	Is the decision linked to addressable patient facts, with patient evidence distinguished from memory, tools, clinical knowledge, and model priors?
Review and revision accountability	Can the reviewer reconstruct the finding, affected target, basis, disposition, responsible stage, and committed change or explicit retention?
Safety and uncertainty transparency	Are risks, assumptions, missing information, dispositions, and unresolved concerns visible through the terminal report?
Trace–output faithfulness	Do the terminal set and explanation preserve recorded decisions, committed revisions, and unresolved risks without introducing unsupported content?
Audit task performance: LLM evaluation	
Audit utility	Can an external reviewer locate a specified medication decision, set change, or safety concern and detect, classify, and localize a controlled trace defect?

Table 22: Inter-rater agreement for the five semantic auditability dimensions.

Comparison	Weighted Cohen's $\kappa\uparrow$
Physician–Physician	0.83
Mean Physician–LLM	0.77

The five semantic dimensions were independently assessed by two physicians, each with more than five years of clinical experience, and by the LLM judge. The LLM judge uses the common model configuration specified in Section 4. Its responses are schema-validated, and cited JSON pointers must resolve within the submitted dossier (Liu et al., 2023; Croxford et al., 2025). Physician ratings were not adjudicated, so disagreements are retained in the reported inter-rater agreement. Weighted Cohen's κ is computed over paired ordinal judgments (Cohen, 1968) and summarized in Table 22. Physician–LLM agreement is the mean of the two κ values obtained by pairing each physician with the LLM judge. These estimates support the cross-rater consistency observed in Figure 2. This retrospective evaluation measures trace quality rather than prescribing appropriateness or prospective clinical benefit in deployment.

H Implementation and Computational Cost

H.1 Implementation Details

EvoMedAgent executes each visit as a bounded sequence of reasoning stages. Clinical problem units are processed independently and may run in parallel, after which their decisions are reconciled at the patient level. Review, optional correction, and faithful summarization then operate on the reconciled result. All LLM-based stages use the common backbone configuration specified in Section 4, and Table 23 reports the principal budgets used in our experiments.

The two learned auxiliary tools are configured separately for MIMIC-III and MIMIC-IV. For each dataset, their estimators are fitted on the training split, while configurations and decision thresholds are selected on the validation split. The resulting tools are invoked only during held-out test inference and therefore do not alter the training-set reasoning traces or failures used for memory evolution.

Table 23: Principal budgets for the auditable reasoning pipeline (Section 3.1) and failure-driven memory evolution (Section 3.2).

Configuration	Value
Auditable Reasoning Pipeline (Section 3.1)	
Concurrent clinical problem units per patient	4
Retrieved experiential entries per entity / unit	16 / 64
Experiential entries exposed per entity / unit prompt	6 / 24
Fallback candidate options per entity / unit prompt	6 / 12
Negative experiential entries per unit prompt	4
Strategic templates per unit	1
Evidence-tool interaction rounds	at most 3
Trace Review–Correction budget	1 correction followed by rereview
Failure-Driven Memory Evolution (Section 3.2)	
Memory-proposal revision budget	at most 1 revision followed by reevaluation
Evolution units per proposal call	at most 4
Prompt characters per proposal call	at most 18,000
Representative FailureCards per evolution unit	at most 10
Experiential support / distinct-case minimum ($n_{\min}$)	2 / 2
Regular experiential confidence threshold (τ)	0.50
Module-observation support for general proposals	at least 5 cases
Test-time memory evolution	disabled

Table 24: API token consumption of one complete EvoMedAgent run on MIMIC-III. Values are in billions of tokens.

Process	Input	Output	Total
Bootstrap initialization	0.042	0.012	0.054
Training-set reasoning pipeline	0.374	0.030	0.404
Memory evolution	0.331	0.028	0.359
Test-set reasoning pipeline	0.111	0.008	0.119
Overall	**0.857**	**0.078**	**0.935**

H.2 Computational Resource Analysis

EvoMedAgent accesses the foundation model through an API rather than training or hosting it locally. This setting avoids the accelerator resources required to optimize or serve the foundation model and permits independent visits and clinical problem units to be processed concurrently. Local computation supports retrieval, the learned auxiliary tools, and deterministic validation. Table 24 reports token consumption across the four stages of one complete MIMIC-III run: bootstrap initialization, training-set reasoning, memory evolution, and test-set inference. Token counts are obtained from the stored requests and completions, with the tokenizer estimates calibrated against provider-reported usage from the corresponding MIMIC-IV calls.

We report token volume and calls per visit rather than wall-clock latency. The experiments execute asynchronous calls through hosted API endpoints, so recorded elapsed time conflates service load, network routing, retries, and worker concurrency with framework computation. Because visits and clinical problem units can be processed concurrently, latency is deployment-dependent, whereas the reported workload measures remain comparable across execution environments.

Training-set reasoning has the largest aggregate input because the full pipeline is executed for every training visit before its trace supports memory evolution. The evolution stage remains comparably expensive because validated failures are compared and grouped before scoped updates are proposed. Test inference consumes more input per visit than training inference because it retrieves from the evolved memory. Bootstrap initialization is comparatively lightweight because it performs label attribution without executing the full reasoning and review workflow.

Table 25: Total LLM API consumption of agent systems on MIMIC-III. Pre-test computation includes memory construction or evolution. Token values combine input and output and are reported in billions. A dash indicates no corresponding stage.

Method	Pre-test tokens (B)	Total test tokens (B)	Test calls / visit
SafeRx-Agent	–	0.121	11.72
SkeMex	0.264	0.057	4.25
EvoMedAgent	0.817	0.119	14.76

Using the published GPT-5.6 Terra rates of $2 per million input tokens and $12 per million output tokens,[2] decision support for one test visit consumes, on average, 0.102 million input tokens and 0.008 million output tokens, corresponding to $0.30. Once outcomes become available, memory evolution amortizes to 0.079 million input tokens and 0.007 million output tokens, or $0.24 per training visit. The latter is a cohort-level average because FailureCards are grouped into shared evolution units before proposals are generated. Under the same price schedule, one complete MIMIC-III run costs approximately $2.6K, while decision support followed by evolution costs approximately $0.54 per visit. These calculations apply standard uncached-input rates.

To contextualize this footprint, Table 25 compares the LLM consumption of EvoMedAgent with two representative agent baselines on MIMIC-III. It separates LLM usage incurred before test inference from test-time token consumption and calls per visit. This is a system-level rather than call-matched comparison: equalizing calls would alter the role structure and native workflow of the evaluated agents, so compute is reported as an explicit system cost.

EvoMedAgent uses a test-time token volume comparable to SafeRx-Agent but distributes its reasoning across more role-specific calls, including unit-level decisions and independent reviews. Its larger pre-test footprint reflects bootstrap initialization, training-set reasoning, and failure-driven memory evolution. SafeRx-Agent has no pre-test memory-evolution stage, while SkeMex incurs a smaller evolution cost.

I Expanded Limitations and Future Directions

Our evaluation examines EvoMedAgent as a retrospective, institution-adaptive medication recommender. The results support its recommendation accuracy, reasoning auditability, and ability to learn reusable experience within this setting. They do not establish autonomous prescribing, causal treatment benefit, or direct transfer of learned memory across clinical environments.

Clinical validity and institutional transfer. MIMIC-III and MIMIC-IV are retrospective datasets from the same healthcare institution. Their coding practices, formulary, patient population, and prescribing patterns consequently shape the candidate associations and experience learned by EvoMedAgent. This institutional adaptation is intentional, since medication vocabularies and treatment practice differ across care settings, but the resulting memory should not be treated as universal clinical knowledge. Historical EHRs may also preserve documentation errors, label noise, or inequitable practice patterns that recurrence-based validation alone cannot identify. Application in a new setting therefore entails reconstructing memory from locally governed data, monitoring performance across patient groups, and prospectively evaluating the resulting system under clinician oversight.

Safety and auditability evidence. EvoMedAgent records DDI, allergy, contraindication, and patient-specific concerns in its review trace, but the quantitative safety analysis is centered on visit-level DDI. An unordered medication set cannot determine concurrent exposure and does not represent dose, route, monitoring, adverse events, or the clinical justification for a necessary high-risk combination. The same limitation applies when DDI is used to screen safety-oriented memory updates. Moreover, the 50-case expert study evaluates whether the reasoning can be traced and reviewed, not whether the recommendation is clinically optimal or improves patient outcomes. Our conclusions therefore concern safety transparency and auditability rather than demonstrated clinical

[2] OpenAI model documentation.

benefit. Time-resolved exposure data and prospective clinical assessment are needed to evaluate that broader outcome.

Memory-evolution evidence. Memory evolution is evaluated against the medication sets observed in the training corpus. Applying each proposed action to all eligible cached predictions tests whether it improves their agreement with observed medications without repeatedly executing the full reasoning pipeline. The resulting gains measure consistency with recorded prescribing practice rather than causal treatment utility, and they do not capture downstream changes that a newly retrieved rule may induce in later agent interactions. The present experiments characterize one complete evolution trajectory, while test repetitions keep the learned memory fixed. They therefore establish the utility of the resulting memory state but not its variability across independent evolution runs. Future evaluations will examine attribution consistency against expert annotations and memory stability across independently evolved runs.

Model and computational scope. The reported system uses one LLM backbone and a fixed set of role-specific tools, so our claims concern the evaluated instantiation of EvoMedAgent. Since tools are integral to each agent's native workflow, the baseline experiments compare complete systems under their respective designs. EvoMedAgent's structured reasoning and corpus-level memory validation require more API computation than simpler agents. Indexed cohort retrieval, cached or batched validation, and smaller models for bounded roles offer natural paths toward greater efficiency without changing the framework's decision and review structure.