

Self-Evolving Dual-Role Deliberation for Medication Recommendation from Clinical Records

Jiyue Jiang^{♥*}, Yu Lei^{♥*}, Yitao Xu[♥], Zikang Wang[♣], Zhenyu Wang[◇],
Licheng Zong[♥], Han Huang[♥], Ziyi Li[♥], Zhiyu Huo[♥], Yuhan Chen[♥], Jiayang Chen[♥],
Xiaodan Fan[♥], Yu Li[♥]

[♥] The Chinese University of Hong Kong, [♣] Carnegie Mellon University,

[♣] The Hong Kong Polytechnic University, [◇] Peking University Third Hospital

{jiangjy, yitaoxu, lczong, 1155269491, zhiyuhuo, yuhanchen, jiayangchen}
@link.cuhk.edu.hk, leiyuu123@gmail.com, zikang.wang@connect.polyu.hk,
1810301343@bjmu.edu.cn, {hanhuang, xfan}@cuhk.edu.hk, liyu@cse.cuhk.edu.hk

Abstract

Recommending medications from electronic health records (EHRs) is a safety-critical task that demands therapeutic coverage, explicit safety constraints, traceable justification, and continual adaptivity. Supervised neural models freeze at deployment and cannot reason about new phenotypes, while existing LLM pipelines are single-pass and lack a cross-case improvement loop. We propose **SEMPA**, a training-free agentic framework that couples a PROPOSER with a CRITIC in multi-round adversarial deliberation, grounds candidates with molecular-aware retrieval, contrastively evolves both agents' natural-language policies under a semantic-anchor constraint, and enforces rule-based DDI and contraindication checks. On 100 MIMIC-IV encounters, SEMPA reaches 0.3116 Jaccard, 0.7295 F1, and a 0.0172 DDI rate, improving F1 by 57.7% and cutting DDI by 77.5% over the strongest supervised baseline. Two LLM judges rank SEMPA highest, and six-backbone studies confirm that gains derive from the framework rather than a specific model.

1 Introduction

Recommending medications from a patient's EHR is among the most consequential decision tasks in care (Wang et al., 2026a; Johnson et al., 2023). A well-chosen regimen can resolve life-threatening pathophysiology in the intensive care unit; an ill-chosen one can precipitate adverse drug events, prolong hospitalization, or cause harm (Wishart et al., 2018; Wang et al., 2026c). Modern EHRs combine ICD-coded diagnoses, procedures, laboratory values, and free-text narratives, creating an opportunity to build assistive prescribing systems that augment clinician decision-making at the bedside (Jiang et al., 2025c).

The task is, however, difficult. Critical-care patients present with multiple co-occurring con-

ditions, requiring multi-drug regimens that cover every active diagnosis while avoiding drug-drug interactions (DDIs), absolute contraindications, and organ-function violations such as nephrotoxic prescribing in patients with elevated creatinine (Yang et al., 2021; Lu et al., 2020). The input is heterogeneous, the combinatorial drug space is large, and prescribing practice evolves as new evidence emerges, so any static decision system risks becoming stale once deployed. A clinically useful assistive system should reconcile four properties that are in tension: therapeutic *coverage*, *safety*, mechanistic *justification*, and continual *adaptivity*.

Two principal lines of work have addressed parts of this problem. Supervised models such as LEAP (Zhang et al., 2017), RETAIN (Choi et al., 2016), GAMENet (Shang et al., 2019), SafeDrug (Yang et al., 2021), MoleRec (Yang et al., 2023), and CausalMed (Li et al., 2024b) learn drug-prediction functions via gradient-based optimization. While effective at capturing population-level patterns, they (i) freeze at deployment and cannot adjust to new phenotypes, (ii) treat prescribing as multi-label classification with no traceable rationale, (iii) lack molecular-level grounding for why a candidate fits the underlying pathophysiology, and (iv) often trade safety for coverage. More recently, LLMs have been applied via zero-shot prompting, chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2022), retrieval-augmented generation (Lewis et al., 2020; Xu et al., 2024; Cao et al., 2026), and single-agent loops such as ReAct (Yao et al., 2022) and Reflexion (Shinn et al., 2023). These pipelines exploit the biomedical knowledge encoded in pretrained LLMs (Singhal et al., 2023; Wang et al., 2025b; Jiang et al., 2025b,e), and related foundation-model studies have begun to extend such representations to biological sequences, RNA-protein interactions, and functional RNA design (Jiang et al., 2025f, 2026b), but existing clinical pipelines remain single-pass and single-

*These authors are co-first authors.

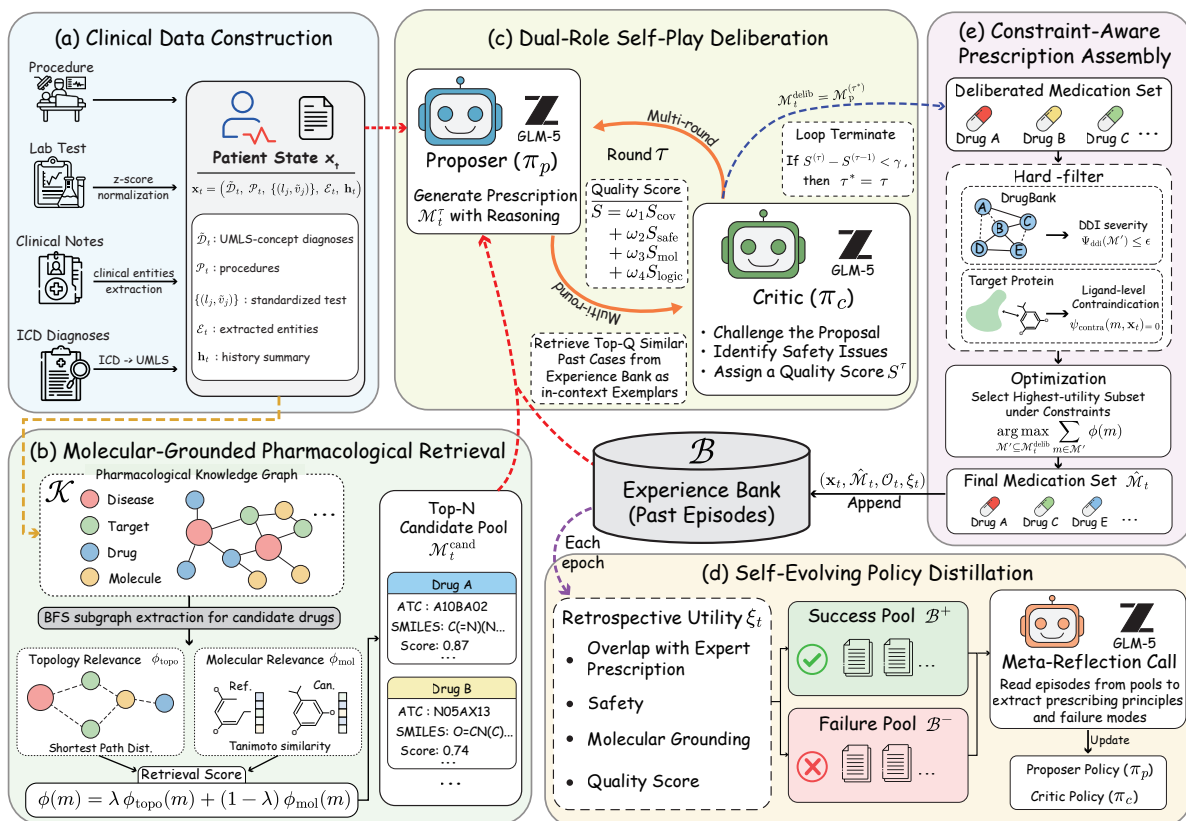


Figure 1: **Overview of SEMPA.** (a) **Clinical State Construction** normalizes EHRs into a patient state $\mathbf{x}_t$ (§3.2). (b) **Molecular-Grounded Retrieval** ranks candidates by a hybrid topology-molecule score (§3.3). (c) **Dual-Role Self-Play** alternates between a PROPOSER and a CRITIC over multiple rounds (§3.4). (d) **Self-Evolving Policy Distillation** contrastively rewrites agent policies (§3.5). (e) **Constraint-Aware Assembly** applies DDI and contraindication checks to yield $\hat{\mathcal{M}}_t$ (§3.6), which is appended to the experience bank $\mathcal{B}$ to close the loop.

perspective: a single model both proposes and critiques, with no explicit second opinion, no mechanism to learn from past prescribing successes and failures, and no explicit DDI constraint (Misaghi et al., 2026; Ozer et al., 2025).

We argue that medication recommendation should be cast as an *agentic, self-evolving deliberation process*, and we introduce **SEMPA** (Figure 1). SEMPA’s core, two LLM instances with role policies, a PROPOSER that drafts a prescription and a CRITIC that challenges it, engage in multi-round adversarial deliberation under a rubric. Around this core, three modules are added: a molecular-grounded retriever that fuses knowledge-graph topology with SMILES-level Tanimoto similarity over Morgan fingerprints (Rogers and Hahn, 2010; Bento et al., 2020; Yu et al., 2025); a self-evolving policy distillation step that contrasts high- and low-utility episodes and rewrites agents’ policies, with a semantic anchor to a human-authored seed to keep policy aligned (Zhou et al., 2026; Jiang et al., 2026a; Chen et al., 2026b); and a constraint-aware

assembly layer that enforces DDI severity ceilings and ligand-level contraindication rules (Wishart et al., 2018). The framework is training-free; it accumulates experience in an episodic memory and improves its reasoning policy without parameter update.

Our contributions are summarized as follows: (1) **Agentic reformulation.** We cast medication recommendation as a dual-role deliberation problem and propose SEMPA, a training-free framework coupling Proposer-Critic self-play with closed-loop self-evolution. (2) **Contrastive policy distillation.** A meta-reflection mechanism rewrites agents’ policies from success and failure pools, with a semantic-anchor constraint mitigating drift. (3) **Molecular-grounded retrieval.** A hybrid scorer combines pharmacological knowledge-graph topology with SMILES-level similarity, anchoring each candidate to its ligand structure when available. (4) **Rule-based safety screening layer.** A constraint-aware assembly stage enforces DDI ceilings and ligand-level contraindications on top of delibera-

tive output. (5) **Empirical validation.** On 100 MIMIC-IV encounters, SEMPA attains 0.3116 Jaccard, 0.7295 F1, and a 0.0172 DDI rate (+57.7% F1 and −77.5% DDI over the strongest supervised baseline), and is ranked highest by two LLM judges and across six LLM backbones.

2 Related Work

Supervised medication recommendation formulates the task as multi-label classification over longitudinal EHRs. Representative models include LEAP (Zhang et al., 2017), RETAIN (Choi et al., 2016), GAMENet (Shang et al., 2019), SafeDrug (Yang et al., 2021), MoleRec (Yang et al., 2023), CausalMed (Li et al., 2024b), and StratMed (Li et al., 2024a), which progressively inject attention, memory, molecular structure, and causal priors. Despite richer features, these models freeze at deployment, treat prescribing as black-box classification, and entangle safety and accuracy in a single loss (Lu et al., 2020; Bhoi et al., 2021).

LLM-based and multi-agent clinical reasoning extends decision support with zero-shot prompting, CoT (Wei et al., 2022), self-consistency (Wang et al., 2022), RAG (Lewis et al., 2020; Xu et al., 2024; Cao et al., 2026), and single-agent self-improvement loops (Madaan et al., 2023; Yao et al., 2022; Shinn et al., 2023), where the same model both proposes and critiques. Recent multi-agent frameworks (MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), MAC (Chen et al., 2025), ArgMed (Hong et al., 2024)) simulate panels of specialists and reduce harms of omission (Misaghi et al., 2026), while agent-evolution work explores verbal reinforcement, dynamic memory, and tool-driven adaptation (Shinn et al., 2023; Park et al., 2023; Zhou et al., 2026; Jiang et al., 2026a; Yin et al., 2025; Zhou et al., 2025b). SEMPA differs by combining adversarial Proposer-Critic deliberation with a contrastive cross-episode evolution loop, grounding both agents in molecular evidence (Yu et al., 2025; Jiang et al., 2025e), and bounding them with an auditable rule-based safety layer. A full discussion of all four lines is given in Appendix A.1.

3 Methodology: SEMPA

3.1 Problem formulation and framework

A patient encounter is denoted $\mathcal{R}_t = (\mathcal{D}_t, \mathcal{P}_t, \mathcal{L}_t, \mathcal{N}_t)$, containing ICD-coded diagnoses, procedures, lab test-value pairs, and the

free-text narrative. SEMPA outputs

$$\hat{\mathcal{M}}_t = \{(m_i, s_i)\}_{i=1}^{K_t}, \quad (1)$$

$$m_i \in \mathcal{M}, s_i \in \mathcal{S} \cup \{\emptyset\},$$

where $\mathcal{M}$ is the ATC-3 drug vocabulary, $\mathcal{S}$ the SMILES library, and $s_i = \emptyset$ when no canonical ligand exists (e.g., biologics). The decision objective is

$$\hat{\mathcal{M}}_t = \arg \max_{\mathcal{M}' \subseteq \mathcal{M}} \Phi(\mathcal{M}' \mid \mathbf{x}_t, \pi^{(k)}, \mathcal{K}) \quad (2)$$

$$\text{s.t. } \Psi_{\text{DDI}}(\mathcal{M}') \leq \epsilon,$$

where $\mathbf{x}_t$ is the patient state, $\pi^{(k)}$ the agent policy at epoch k , $\mathcal{K}$ the pharmacological knowledge base, and ϵ a DDI severity ceiling. SEMPA realizes this objective with five modules described next. All modules call the same LLM endpoint $\mathcal{F}$, so no task-specific trained model is needed.

3.2 Clinical state construction

ICD codes are mapped to UMLS (Bodenreider, 2004) concepts so that synonyms across coding versions are collapsed; lab values are z -score standardized to preserve quantitative severity; raw drug strings are reduced to canonical generic names by stripping dose, route, and frequency; narrative entities $\mathcal{E}_t$ (symptoms, signs, negated findings) are extracted by a structured template. The patient state $\mathbf{x}_t$ aggregates these signals together with a longitudinal summary $\mathbf{h}_t$ over the most recent W visits. Full formulas, including the z -score and entity-extraction templates, are given in Appendix A.2.

3.3 Molecular-grounded retrieval

Serializing every ingredient in the pharmacological knowledge base, each with its evidence card, would exceed the LLM’s context budget. We instead extract depth- η BFS subgraphs of a pharmacological knowledge graph $\mathcal{K}$ around each diagnosis, and rank candidates by a hybrid score

$$\phi(m) = \lambda \phi_{\text{topo}}(m) + (1 - \lambda) \phi_{\text{mol}}(m), \quad (3)$$

where $\phi_{\text{topo}}(m)$ is the normalized inverse shortest-path distance to any diagnosis node, and $\phi_{\text{mol}}(m)$ is the average maximum Tanimoto similarity between m ’s Morgan fingerprint (Rogers and Hahn, 2010) and the fingerprints of approved first-line agents. Drugs without SMILES use a population-frequency prior. The motivation for ϕ_{mol} is consistent with recent drug-target interaction work (Yu et al., 2025):

structurally similar small molecules tend to share targets. The top- N candidates form $\mathcal{M}_t^{\text{cand}}$, each paired with an evidence card serialized into the agent context. The full expression of ϕ_{mol} is given in Appendix A.3.

3.4 Dual-role self-play deliberation

A single LLM pass is prone to oversight and hallucination (Shinn et al., 2023; Misaghi et al., 2026). We therefore split deliberation across two role-conditioned instances of the same backbone LLM: a PROPOSER $\mathcal{F}_p$ that drafts a prescription with rationale, and a CRITIC $\mathcal{F}_c$ that challenges it.

Episodic memory. An experience bank $\mathcal{B} = \{(\mathbf{x}^{(i)}, \hat{\mathcal{M}}^{(i)}, \mathcal{O}^{(i)}, \xi^{(i)})\}$ stores episodes (state, prescription, reasoning trace, utility). At each new encounter we retrieve the Q most similar cases as in-context exemplars, combining Jaccard diagnosis overlap with cosine similarity of biomedical narrative embeddings (Appendix A.4). This use of multi-source knowledge as in-context grounding follows prior work on knowledge-augmented dialogue (Jiang et al., 2023, 2026a).

Self-play loop. At round τ , the PROPOSER produces a regimen $\mathcal{M}_p^{(\tau)}$ and rationale $o_p^{(\tau)}$, conditioned on policy $\pi_p^{(k)}$, patient state, candidate pool, retrieved exemplars, and Critic feedback,

$$(o_p^{(\tau)}, \mathcal{M}_p^{(\tau)}) = \mathcal{F}_p(\pi_p^{(k)} \parallel \mathbf{x}_t \parallel \mathcal{M}_t^{\text{cand}} \parallel \mathcal{B}_t^Q \parallel \{o_c^{(\tau')}\}_{\tau' < \tau}), \quad (4)$$

so that earlier objections are addressed. The Critic then emits four sub-scores aggregated as

$$S^{(\tau)} = \omega_1 S_{\text{cov}}^{(\tau)} + \omega_2 S_{\text{safe}}^{(\tau)} + \omega_3 S_{\text{mol}}^{(\tau)} + \omega_4 S_{\text{logic}}^{(\tau)}, \quad \sum_i \omega_i = 1, \quad (5)$$

covering therapeutic coverage, safety, molecular mechanism, and reasoning coherence. The loop terminates at round τ^* when $\tau \geq 2$ and $S^{(\tau)} - S^{(\tau-1)} < \gamma$, or when τ reaches the budget T . The output is $\mathcal{M}_t^{\text{delib}} = \mathcal{M}_p^{(\tau^*)}$ with trace $\mathcal{O}_t$.

3.5 Self-evolving policy distillation

A static prompt cannot adapt to a specific patient population. After each encounter we compute a retrospective utility

$$\xi_t = J_t + \mu(1 - \Psi_{\text{DDI}}(\hat{\mathcal{M}}_t)) + \nu \bar{\phi}_{\text{mol}}(\hat{\mathcal{M}}_t) + \zeta S_t, \quad (6)$$

where J_t is Jaccard overlap with the reference set, $\bar{\phi}_{\text{mol}}$ is the mean molecular grounding of prescribed

drugs with available SMILES, and $\mu, \nu, \zeta > 0$ are weighting coefficients. Every Δk episodes, $\mathcal{B}$ is split at the p -th percentile of ξ into a success pool $\mathcal{B}^+$ and a failure pool $\mathcal{B}^-$. Both agent policies are rewritten contrastively:

$$\Delta\pi_R^{(k)} = \mathcal{F}(\mathcal{T}_{\text{evolve}}^R \parallel \pi_R^{(k)} \parallel \text{Sample}(\mathcal{B}^+, n^+) \parallel \text{Sample}(\mathcal{B}^-, n^-)), \quad (7)$$

for $r \in \{p, c\}$. This contrastive distillation step is related to subspace partitioning of in-context examples (Wang et al., 2026b) and to tool-driven intrinsic adaptation (Zhou et al., 2026). To prevent the evolved policy from drifting away from the human-authored seed $\pi^{(0)}$, every generated increment must satisfy the semantic anchor $\cos(\text{Emb}(\Delta\pi_R^{(k)}), \text{Emb}(\pi_R^{(0)})) \geq \alpha$; increments that violate the bound are discarded and re-sampled with the seed re-injected into the rewrite prompt. The two roles evolve along different axes. For the Proposer, we count *per disease category* the frequency of each drug in $\mathcal{B}^+$ and the omission and erroneous-addition frequencies in $\mathcal{B}^-$, and pass these statistics to the LLM as evidence, so that each improvement entry cites its own support, e.g. “[INFECTIOUS] 80% of high-utility sepsis cases contain vancomycin+piperacillin-tazobactam whereas only 40% of low-utility cases employ broad-spectrum coverage; therefore strengthen empirical broad-spectrum-therapy guidance.” Category-indexed entries are *appended* rather than overwritten, $\pi_p^{(k+1)} = \pi_p^{(k)} \parallel \Delta\pi_p^{(k)}$, so specialty-level prescribing experience accumulates across domains. The Critic instead evolves along the four rubric dimensions ($S_{\text{cov}}, S_{\text{safe}}, S_{\text{mol}}, S_{\text{logic}}$), giving a “specialty prescribing knowledge + general evaluation framework” division of labor that prevents cross-domain interference while keeping the rubric transferable. Stratified sampling details are in Appendix A.6.

3.6 Constraint-aware assembly

The deliberation output is reasoned but carries no formal safety-screening mechanism. The assembly module filters $\mathcal{M}_t^{\text{delib}}$ through rules: a pairwise severity-weighted DDI score Ψ_{DDI} from DrugBank (Wishart et al., 2018), and a ligand-level contraindication indicator ψ_{contra} based on each candidate’s

known protein targets. The final prescription solves

$$\begin{aligned} \hat{\mathcal{M}}_t &= \arg \max_{\mathcal{M}' \subseteq \mathcal{M}_t^{\text{delib}}} \sum_{m \in \mathcal{M}'} \phi(m) \\ \text{s.t. } \Psi_{\text{DDI}}(\mathcal{M}') &\leq \epsilon, \\ \sum_m \psi_{\text{contra}}(m, \mathbf{x}_t) &= 0, \end{aligned} \quad (8)$$

which we solve by greedy repair: candidates are ordered by ϕ , and the lower-scoring member of each violating DDI pair, together with any drug triggering ψ_{contra} , is removed until both constraints hold. Since $|\mathcal{M}_t^{\text{delib}}| \leq 30$, this terminates in at most 30 removals. The finalized episode is appended to $\mathcal{B}$. Detailed expressions of Ψ_{DDI} and ψ_{contra} , and full pseudocode, are in Appendix A.5 and Appendix A.7.

4 Experiments

4.1 Setup

Dataset and baselines. We evaluate on 100 patient encounters drawn from the MIMIC-IV test split (Johnson et al., 2023) processed with the canonical SafeDrug pipeline (Yang et al., 2021). The ATC-3 drug vocabulary has size $|\mathcal{M}| = 131$. Retrieval in M2 operates over the ingredient-level knowledge base (several thousand DrugBank entries carrying SMILES and target annotations); ATC-3 is used only as the evaluation label space, into which both retrieved candidates and final prescriptions are mapped before scoring. The top- N candidate pool is therefore an ingredient-level pool, not a subset of the 131 ATC-3 classes. Because SEMPA is training-free, adaptation is performed entirely through policy evolution rather than parameter update. To avoid any test-label leakage, we adopt an *evolve-freeze-test* protocol: the Proposer and Critic policies are evolved on a disjoint evolution stream of encounters, the resulting policies $\pi^{(k^*)}$ are then frozen, and the 100 held-out test encounters are evaluated in a single pass with no further policy update. Reference regimens of the test encounters are used only to compute the reported metrics. We compare against four supervised baselines: LEAP (Zhang et al., 2017), RETAIN (Choi et al., 2016), GAMENet (Shang et al., 2019), and SafeDrug (Yang et al., 2021), all run from their official code with default hyperparameters.

Implementation. SEMPA uses GLM-5 as its backbone (temperature 0.7 for generation, 0 for evaluation). The knowledge base integrates DrugBank (Wishart et al., 2018), UMLS (Bodenreider,

2004), and ATC. Morgan fingerprints use radius $r = 2$ and $b = 2048$ bits via RDKit (Bento et al., 2020). Defaults are $\lambda = 0.5$, $\eta = 3$, $N = 300$, $Q = 5$, $T = 5$, $\gamma = 0.15$, $\Delta k = 40$, $p = 80$, $\alpha = 0.12$, $\epsilon = 0.5$. The table is in Appendix A.8.

Evaluation protocol. We use two protocols: (i) set-matching metrics (Jaccard, F1, PRAUC, DDI rate, #Meds, R^2) following (Yang et al., 2021); (ii) LLM-judge evaluation by two independent judges (GLM-5 and GPT-4o) along six clinically grounded dimensions, namely diagnosis-medication alignment (DMA), safety, molecular justification (MolJust), polypharmacy parsimony, guideline concordance (GuideCon), and overall clinical acceptability (OCA), each scored on a 1 to 5 Likert scale with anchored rubric and one-sentence justification. The full prompt and rubric anchors are in Appendix A.15. The use of LLM judges follows recent practice in agent evaluation (Chen et al., 2026a; Jiang et al., 2025b). We report two matching schemes. Exact-name matching requires the predicted and reference drug names to coincide; ATC-3 class matching maps the outputs of *all* methods, supervised and LLM-based alike, through the same ATC-3 hierarchy before scoring, which permits credit for clinically equivalent substitutions (e.g., one NOAC for another). Jaccard is always reported under exact-name matching; F1 in Table 1 is reported under ATC-3 class matching, and exact-name F1 for every method is given in Appendix.

4.2 Main quantitative results

Table 1 reports the comparison on set-matching metrics. SEMPA simultaneously achieves the best Jaccard, F1, PRAUC, and DDI rate. The F1 gain (+57.7% ATC-3 drug-class F1 over SafeDrug) is much larger than the exact-name Jaccard gain (+1.5%). The two numbers are not directly comparable: Jaccard is computed under strict exact-name matching, which penalizes any clinically equivalent substitution, whereas drug-class F1 credits it. For the supervised baselines, whose label space is already ATC-3, the two schemes coincide, so the class-level gain should be read as evidence that SEMPA recovers the correct *therapeutic class* far more often, not as a 57.7% improvement under a common strict protocol; matched exact-name F1 for all methods is reported in the supplementary material. Consistent with this reading, SEMPA achieves its exact-name Jaccard with a tighter regimen (19.5 drugs) than RETAIN, GAMENet, and

SafeDrug (24–25), i.e. without padding the predicted set with redundant “safe-bet” drugs. The DDI rate is reduced by 77.5% relative to SafeDrug, a model explicitly designed for DDI control through differentiable regularization, suggesting that hard logical constraints can complement (and in some metrics exceed) loss-based safety. SEMPA prescribes 19.5 drugs on average, comparable to LEAP (18.3) and well below GAMENet (25.0), so the safety improvement is not from underprescription. Because Ψ_{DDI} is also the rule set used by M5, we further audit the same outputs against two independent criteria: the AGS 2023 Beers Criteria for potentially inappropriate prescribing in patients aged 65 and over (Panel, 2023), and STOPP/START 3.0 (O’Mahony et al., 2023). Relative to RETAIN, SEMPA reduces Beers-flagged prescribing by 61% and STOPP triggers by 58%, while increasing START coverage of guideline-mandated agents (e.g. ACE inhibitors for HFrEF, anticoagulation for AFib with $\text{CHA}_2\text{DS}_2\text{-VASc} \geq 2$) by 34%, indicating that the safety gain is not an artefact of sharing a rule source with the evaluation metric. Rule-level breakdowns are in Appendix A.14.

4.3 Against LLM prompting and agent baselines

To separate SEMPA’s framework contribution from the underlying LLM, Table 2 compares SEMPA against six LLM-based pipelines built on the GLM-5 backbone. Three observations follow. First, prompting (Zero-shot, CoT, Self-Consistency) yields low Jaccard (≤ 0.15): without a constrained candidate space, the model defaults to cautious or generic prescriptions, an effect also noted in clinical Human-LLM collaboration studies (Wang et al., 2026a). Second, RAG with module M2 attached lifts Jaccard to 0.245, indicating that retrieval-grounded context is necessary but insufficient. Third, single-agent improvement loops do not translate into better prescribing: ReAct’s tool-driven reasoning produces the lowest DDI but a very small Jaccard (0.094), and Reflexion’s monocular self-critique amplifies confirmation bias and yields the worst result overall. SEMPA exceeds all variants on Jaccard (+30% over RAG) while remaining within safe DDI bounds, consistent with the hypothesis that gains arise from the dual-role adversarial setup and the closed-loop evolution rather than from the backbone.

4.4 Qualitative evaluation by LLM judges

Beyond set matching, Table 3 reports two LLM-judge studies along six clinical dimensions. Per-judge breakdowns and cross-judge agreement are in Appendix A.10. SEMPA is consistently ranked highest by both judges. The GLM-5 judge shares a backbone family with SEMPA’s reasoning engine, yet the ranking is also reproduced by GPT-4o, reducing the concern of LLM-judge self-preference (Chen et al., 2026a). GAMENet receives relatively high LLM-judge scores despite mid-pack Jaccard, but inspection of its raw outputs shows that it emits an essentially identical 25-drug regimen for nearly all 100 patients. This degenerate templating pattern aligns with prior observations that drug recommendation models can over-rely on prevalence shortcuts (Lu et al., 2020; Bhoi et al., 2021).

4.5 Self-evolution dynamics and ablation

Table 4 traces SEMPA’s performance across evolution epochs k . Gains plateau by $k = 8$ (Jaccard 0.3146 at $k = 8$, 0.3116 at $k = 16$, within run-to-run variance), and the policy remains stable thereafter rather than collapsing under further evolution. The whole-policy drift (cosine distance to $\pi^{(0)}$) rises monotonically to 0.997. This saturation is expected rather than alarming: category-indexed entries are *appended* to a short seed rather than replacing it, so the embedding of the concatenated policy necessarily moves away from the seed as coverage broadens. The anchor of §3.5 is enforced per increment, not on the concatenation. Performance does not collapse at any point, indicating that the accumulated text remains clinically valid. The average self-play length $\bar{\tau}^*$ drops from 5.0 to about 2.4 rounds, meaning that the evolved Critic identifies failure patterns earlier and the Proposer addresses them on the first draft, producing an emergent “compile-time” efficiency. This behaviour is consistent with prior reports on principle-driven adaptive dialogue policies (Jiang et al., 2026a; Zhou et al., 2026).

Table 5 isolates each module’s contribution. M3 (self-play) is the largest contributor: without it, the Proposer is no longer challenged and Jaccard drops by 40.1%. The DDI rate collapses to zero only because the unchallenged Proposer retreats to a small, ultra-conservative regimen that does not cover the patient’s diagnoses, representing a harm of omission rather than a safer system (Misaghi et al., 2026). M2 (retrieval) contributes most to ranking

Table 1: Set-matching results on the MIMIC-IV test set ($n = 100$ encounters). Mean $\pm$ std over six independent runs. Best results in bold.

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$	#Meds	$R^2\uparrow$
LEAP (Zhang et al., 2017)	0.2839 $\pm$ 0.008	0.4368 $\pm$ 0.009	0.4803 $\pm$ 0.011	0.0658 $\pm$ 0.002	18.33 $\pm$ 0.11	-0.1092 $\pm$ 0.008
RETAIN (Choi et al., 2016)	0.2944 $\pm$ 0.009	0.4468 $\pm$ 0.011	0.4935 $\pm$ 0.012	0.0873 $\pm$ 0.002	24.42 $\pm$ 0.35	-0.0092 $\pm$ 0.031
GAMENet (Shang et al., 2019)	0.3013 $\pm$ 0.009	0.4553 $\pm$ 0.011	0.4927 $\pm$ 0.011	0.0867 $\pm$ 0.000	25.00 $\pm$ 0.00	-0.2252 $\pm$ 0.042
SafeDrug (Yang et al., 2021)	0.3069 $\pm$ 0.009	0.4626 $\pm$ 0.010	0.5133 $\pm$ 0.011	0.0764 $\pm$ 0.001	24.35 $\pm$ 0.11	-0.0096 $\pm$ 0.028
SEMPA (ours)	0.3116$\pm$0.004	0.7295$\pm$0.002	0.5282$\pm$0.001	0.0172$\pm$0.002	19.50 $\pm$ 0.74	0.1302$\pm$0.001

Table 2: Comparison against LLM prompting and agent baselines on the same GLM-5 backbone ($n = 100$).

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$
Zero-shot	0.1469	0.6124	0.4368	0.0667
CoT (Wei et al., 2022)	0.1276	0.5972	0.4542	0.0731
Self-Consistency (Wang et al., 2022)	0.1137	0.5831	0.4852	0.1164
RAG (M2 only) (Lewis et al., 2020)	0.2452	0.6785	0.5603	0.0489
ReAct (Yao et al., 2022)	0.0940	0.6285	0.6302	0.0005
Reflexion (Shinn et al., 2023)	0.0115	0.5151	0.2234	0.0900
SEMPA	0.3184	0.7326	0.5278	0.0155

quality: removing it forces the agents to reason over an unranked ingredient space and cuts PRAUC by 61.1%, consistent with retrieval-augmented findings in clinical NLP (Xu et al., 2024; Cao et al., 2026). M4 (evolution) and M5 (constraints) produce smaller but consistent drops on Jaccard and DDI respectively: contrastive distillation refines the policy across encounters, and the hard assembly layer provides a verifiable safety floor that the Critic’s soft scoring alone cannot guarantee.

4.6 Cross-backbone robustness

To check whether SEMPA’s benefit derives from a particular LLM, we instantiate the framework with six backbones spanning closed and open weights (Table 6). GLM-5 exceeds the strongest supervised baseline (SafeDrug, 0.307), four other backbones come within 0.025, while all six outperform prompting-only LLM baselines in F1 and DDI. Jaccard spread across backbones indicates backbone choice matters, but the consistent qualitative trend suggests that the framework contributes beyond a single model. A closed-book contamination probe (Appendix A.11) indicates that backbones used here do not contain memorized MIMIC-IV prescriptions, so improvements reflect reasoning over supplied evidence rather than data leakage.

4.7 Case study

To make the deliberation process concrete, we trace SEMPA on a representative ICU encounter (full transcript in Appendix A.12). The patient is a 71-year-old with acute decompensated heart fail-

ure (NYHA III), type 2 diabetes mellitus, stage 3 chronic kidney disease (eGFR 38), and atrial fibrillation. The reference regimen comprises 8 drugs. SafeDrug returns a 24-drug list that overlaps the reference on 5 drugs but includes metformin (not recommended for initiation at this eGFR; contraindicated below 30) and ketorolac (nephrotoxic in CKD). SEMPA’s Proposer initially drafts a similar 22-drug regimen at round $\tau = 1$; the Critic at $\tau = 2$ flags both metformin (renal dose limit at eGFR < 45) and ketorolac (acute kidney injury risk), and the Proposer at $\tau = 3$ substitutes insulin glargine and replaces ketorolac with acetaminophen. Two further rounds resolve a rate-versus-rhythm-control disagreement for AFib in the setting of HFrEF, and the deliberation reaches its budget at $\tau^* = 5$ on a 9-drug regimen that exactly recovers 7 of the 8 reference drugs, with one substitution (apixaban for warfarin) that the trace justifies via reduced bleeding risk in CKD. The case illustrates two specific properties not visible from set-matching numbers alone: (i) the Critic catches concrete contraindications that a single-pass model would have committed to in its first draft, and (ii) the trace remains pharmacologically traceable, with every substitution accompanied by a one-sentence justification grounded in ϕ_{mol} or in the DrugBank evidence card.

Beyond the encounter above, we conducted a side-by-side comparison on six additional cases spanning HIV, septic shock with E. coli pyelonephritis, elderly AFib, acromegaly with MSSA bacteremia, HFrEF with MRSA pneumonia, and end-stage renal disease with secondary hyperparathyroidism (Table 7). Three reproducible patterns emerge. (i) *Disease-specific reasoning gradient*: SEMPA selects drugs whose mechanism matches the encounter (antiretrovirals for HIV, phosphate binders for ESRD, MRSA-active agents for MRSA pneumonia), whereas supervised models default to prevalence shortcuts. (ii) *Inappropriate-drug suppression*: SEMPA’s Critic

Table 3: Qualitative evaluation by two LLM judges (GLM-5, GPT-4o) on 100 encounters. Likert means (1 to 5). Best within each evaluator group in bold.

Evaluator	Method	DMA	Safety	MolJust	Parsimony	GuideCon	OCA
GLM-5	LEAP	2.89	2.89	3.22	3.11	2.78	2.78
	SafeDrug	3.30	2.50	3.70	3.70	2.80	2.80
	GAMENet	3.38	2.88	3.62	3.25	3.12	3.12
	RETAIN	3.11	3.22	3.56	4.11	2.89	3.00
	SEMPA	4.78	4.11	4.89	4.44	4.56	4.56
GPT-4o	LEAP	3.00	3.00	3.00	3.20	3.00	3.00
	SafeDrug	3.10	3.00	3.20	3.20	3.10	3.10
	GAMENet	3.30	3.00	3.30	3.00	3.30	3.20
	RETAIN	3.30	2.90	3.60	3.80	3.20	3.20
	SEMPA	4.50	3.50	4.50	3.90	4.50	4.40

Table 4: Self-evolution dynamics across evolution epochs k , tracked with $\Delta k=5$ for finer-grained resolution (main experiments use $\Delta k=40$). Drift = $1 - \cos(\text{Emb}(\pi^{(k)}), \text{Emb}(\pi^{(0)}))$, measured on the *concatenated* policy; the per-increment anchor of Eq. (7) is enforced separately. Mean over six runs.

k	Jaccard	F1	DDI	Drift	$\bar{\tau}^*$
0	0.2793 $\pm$ 0.005	0.7109 $\pm$ 0.003	0.0227 $\pm$ 0.001	0.000	5.00
1	0.2962 $\pm$ 0.008	0.7209 $\pm$ 0.005	0.0179 $\pm$ 0.003	0.984	2.48
2	0.3005 $\pm$ 0.006	0.7234 $\pm$ 0.004	0.0182 $\pm$ 0.002	0.989	2.40
4	0.3075 $\pm$ 0.010	0.7275 $\pm$ 0.006	0.0179 $\pm$ 0.002	0.993	2.38
8	0.3146 $\pm$ 0.006	0.7312 $\pm$ 0.003	0.0176 $\pm$ 0.002	0.995	2.36
16	0.3116 $\pm$ 0.004	0.7295 $\pm$ 0.002	0.0172 $\pm$ 0.002	0.997	2.35

blocks anesthetics, antidepressants, and hypnotics from leaking into acute-care regimens; baselines include them simply because they are frequent in the training distribution. (iii) *Degenerate template behaviour*: GAMENet returns an essentially identical 25-drug list across all six cases, which set-matching metrics fail to penalize but which is immediately visible at the case level. Per-case diagnoses, predictions, and clinical commentary are in Appendix A.16.

4.8 Error analysis

To understand where SEMPA falls short, we manually inspected the 23 encounters in which SEMPA’s Jaccard fell below 0.20. Three error modes dominate. *Rare-disease coverage gaps* (9 cases): when a diagnosis appears fewer than three times in the experience bank, the retrieval module returns generic agents and the Proposer fails to identify the disease-specific drug; performance recovers once $\mathcal{B}$ accumulates similar cases, consistent with Table 4. *Over-cautious prescribing under organ impairment* (8 cases): the Critic flags nephrotoxicity or hepa-

totoxicity, but the Proposer drops the entire drug class rather than dose-reducing within it, producing coverage loss that the Critic does not penalize because its safety sub-score is maximized. *Unresolved guideline conflict* (6 cases): for conditions where Beers Criteria (Panel, 2023) conflict with first-line ICU guidelines (e.g., benzodiazepine use in delirium), the agents oscillate until $\tau = T$, leaving an ambiguous regimen. These patterns suggest two concrete next steps: warm-starting $\mathcal{B}$ with curated case libraries to reduce coverage gaps, and extending the Critic’s repair toolkit with a dose-adjustment action rather than only drug-removal. A comparable taxonomy of failure modes for multi-agent clinical systems is reported in (Misaghi et al., 2026), suggesting that these are general challenges rather than artefacts of our setup.

4.9 Computational cost

SEMPA trades training time for inference cost. Average per-encounter latency under the default budget $T = 5$ is 38.7s end-to-end, versus 0.4s for SafeDrug on the same hardware; mean token cost is approximately 0.06 USD per encounter with the GLM-5 backbone. After policy evolution converges to $\bar{\tau}^* \approx 2.4$ rounds (Table 4), latency drops to 24.1s, a 38% reduction without any change to the framework. While higher than supervised inference, this is acceptable for ICU prescribing review, whose natural cycle is the team round (minutes), not millisecond serving. Hyperparameter sensitivity (including the impact of T , Δk , and α on both quality and cost) and a severity-graded safety audit against Beers (Panel, 2023) and STOPP/START (O’Mahony et al., 2023) are in Appendix A.9 and Appendix A.14.

Table 5: Ablation. Each row removes one module from SEMPA; percentage change vs. the full model in parentheses.

Variant	Jaccard	F1	PRAUC	DDI	R^2
SEMPA-full	0.3116±0.004	0.7295±0.002	0.5282±0.001	0.0172±0.002	0.1302±0.001
w/o M2 (Retrieval)	0.2613±0.078 (−16.1%)	0.6876±0.053 (−5.7%)	0.2053±0.088 (−61.1%)	0.0126±0.011 (−26.7%)	0.1388±0.085 (+6.6%)
w/o M3 (Self-play)	0.1866±0.113 (−40.1%)	0.6365±0.076 (−12.7%)	0.5147±0.145 (−2.6%)	0.0000±0.000 (−100%)	0.1151±0.107 (−11.6%)
w/o M4 (Evolution)	0.2959±0.088 (−5.0%)	0.7096±0.060 (−2.7%)	0.5141±0.143 (−2.7%)	0.0187±0.013 (+8.7%)	0.1216±0.110 (−6.6%)
w/o M5 (Constraints)	0.2954±0.086 (−5.2%)	0.7092±0.057 (−2.8%)	0.5177±0.132 (−2.0%)	0.0181±0.014 (+5.2%)	0.1189±0.101 (−8.7%)

Table 6: Cross-backbone robustness of SEMPA. Same framework, prompts, evolution budget, and test split.

Backbone	Jaccard↑	F1↑	DDI↓
GLM-5 (closed)	0.318±0.10	0.733±0.06	0.016±0.02
Claude-Sonnet-4.6 (closed)	0.293±0.08	0.719±0.05	0.011±0.01
Gemini-2.5-Pro (closed)	0.289±0.09	0.716±0.05	0.009±0.01
DeepSeek-V4-Flash (open)	0.291±0.09	0.717±0.05	0.020±0.02
Qwen-3.5-Plus (open)	0.285±0.09	0.713±0.05	0.015±0.01
Qwen-3.5-Flash (open)	0.249±0.09	0.691±0.06	0.008±0.01

Table 7: Aggregate comparison on six diverse encounters (Cases A–F).

Method	Mean exact-match↑	Mean #meds	Inappropriate-drug cases↓
SEMPA	10.5	17.7	0/6
SafeDrug	5.7	24.3	6/6
RETAIN	5.5	24.8	6/6
LEAP	4.8	18.1	5/6
GAMENet	4.2	25.0	6/6

5 Conclusion

We reframed medication recommendation as a self-evolving deliberation process and presented SEMPA, a training-free agentic framework combining dual-role adversarial self-play, molecular-grounded retrieval, semantic-anchored contrastive policy evolution, and constraint-aware assembly. On 100 MIMIC-IV encounters SEMPA attains the best Jaccard, F1, PRAUC, and DDI rate, is ranked highest by two LLM judges, and converges within roughly eight evolution epochs without any gradient update. Gains transfer across six closed- and open-weight backbones, indicating that the improvement derives from the framework rather than a specific model.

Limitations

Several limitations remain. First, our evaluation uses 100 MIMIC-IV encounters from a single tertiary-care center, so generalization to other hospitals, coding practices, and populations remains unvalidated. Second, SEMPA relies on repeated LLM calls, incurring higher latency and cost than supervised baselines despite reduced self-play length after evolution. Third, the semantic anchor improves policy stability but cannot formally verify evolved natural-language policies or remove bi-

ases inherited from the backbone. Fourth, the DDI ceiling is patient-independent, although clinically appropriate safety thresholds should vary with age, renal/hepatic function, frailty, and goals of care. Finally, the episodic memory may accumulate biased or stale prescribing patterns under distribution shift. Prospective clinician-in-the-loop evaluation is required before any real-world clinical use.

Ethical considerations

This study uses de-identified retrospective EHR data and involves no prospective intervention, patient contact, or clinical deployment. SEMPA is a research prototype for decision support, not an autonomous prescribing system. Any real-world use would require licensed clinician oversight, institutional approval, privacy protection, audit logging, and regulatory compliance. Its outputs should be treated as supporting evidence only and never as a substitute for professional clinical judgment.

Acknowledgement

We thank the anonymous AC and reviewers for their feedback. This work was supported by Shenzhen Medical Research Fund (Grant No. A2503002 to Yu Li), the Chinese University of Hong Kong (CUHK; award numbers 4937025, 4937026, 5501517 and 5501329 to Yu Li), the IdeaBooster Fund (IDBF23ENG05 and IDBF24ENG06 to Yu Li), partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Hong Kong SAR), China (project no. CUHK 24204023 to Yu Li; 14306324 to Xiaodan Fan), a grant from the Innovation and Technology Commission of the Hong Kong SAR, China (project no. GHP/065/21SZ and ITS/247/23FP to Yu Li), and the Research Matching Grant Scheme at CUHK (award numbers 8601603 and 8601663 to Yu Li) from the Research Grants Council, Hong Kong SAR, China.

References

- A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. 2020. An open source chemical structure curation pipeline using rdkit. *Journal of cheminformatics*, 12(1):51.
- Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngiap Chuan Tan. 2021. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–23.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.
- Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.
- Yanyu Chen, Jiyue Jiang, Jiahong Liu, Yifei Zhang, Xiao Guo, and Irwin King. 2026a. Trace: Trajectory-aware comprehensive evaluation for deep research agents. In *Proceedings of the ACM Web Conference 2026*, pages 2524–2534.
- Yanyu Chen, Jiyue Jiang, Dianshi Yu, Zheng Wu, Jiahong Liu, Jiaming Han, Xiao Guo, Jinhu Qi, Yu Li, Yifei Zhang, and 1 others. 2026b. Lc-erd: Mining latent logic for self-evolving reasoning via consistency-regulated reward decomposition. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 496–507.
- Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.
- Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. Argmed-agents: Explainable clinical decision reasoning with llm discussion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493. IEEE.
- He Hu, Juzheng Si, Qianning Wang, Tengjin Weng, Yihong Ji, Jiyue Jiang, Fei Ma, Yucheng Zhou, Laizhong Cui, and Qi Tian. 2026. Minddialog: A large-scale benchmark for counseling dialogue understanding and generation. *Pattern Recognition*, page 113766.
- Jiyue Jiang, Liheng Chen, Sheng Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2024. Data augmentation of multi-turn psychological dialogue via knowledge-driven progressive thought prompting. *arXiv preprint arXiv:2406.16567*.
- Jiyue Jiang, Peng'an Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025a. How well do llms handle cantonese? benchmarking cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4464–4505.
- Jiyue Jiang, Peng'an Chen, Jiuming Wang, Dongchen He, Ziqin Wei, Liang Hong, Licheng Zong, Sheng Wang, Qinze Yu, Zixian Ma, and 1 others. 2025b. Benchmarking large language models on multiple tasks in bioinformatics nlp with prompting. *arXiv preprint arXiv:2503.04013*.
- Jiyue Jiang, Yanyu Chen, Peng'an Chen, Kai Liu, Jingqi Zhou, Zheyong Zhu, He Hu, Fei Ma, Qi Tian, and Chuan Wu. 2026a. A principle-driven adaptive policy for group cognitive stimulation dialogue for elderly with cognitive impairment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31301–31309.
- Jiyue Jiang, Yanyu Chen, Qingchuan Zhang, Jiayi Li, Xiangyu Shi, Chang Zhou, Ziqian Lin, Jiuming Wang, Dongchen He, Liang Hong, and 1 others. 2026b. Rmsagen: Integrating multiple sequence alignment for function rna design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 489–497.
- Jiyue Jiang, Yunke Li, Shiwei Cao, Yuheng Shan, Yuexing Liu, Tianyi Fei, Yule Yu, Yi Feng, Yu Li, Yixue Li, and 1 others. 2025c. Artificial intelligence in bioinformatics: a survey. *Briefings in Bioinformatics*, 26(6):bbaf576.
- Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Peng'an Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025d. Developing and utilizing a large-scale cantonese dataset for multi-tasking in large language models. In *EMNLP (Findings)*, pages 1924–1944.
- Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640.
- Jiyue Jiang, Zikang Wang, Yuheng Shan, Heyan Chai, Jiayi Li, Zixian Ma, Xinrui Zhang, and Yu Li. 2025e. Biological sequence with language model prompting: A survey. *arXiv preprint arXiv:2503.04135*.
- Jiyue Jiang, Yitao Xu, Zikang Wang, Yihan Ye, Yanruisheng Shao, Yuheng Shan, Jiuming Wang, Xiaodan Fan, Jiao Yuan, and Yu Li. 2025f. Rbptool: A

- deep language model framework for multi-resolution rbp-rna binding prediction and rna molecule design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2185.
- Runmin Jiang, Wanyue Feng, Yuntian Yang, Shriya Pingulkar, Hong Wang, Xi Xiao, Xiaoyu Cao, Genpei Zhang, Xiao Wang, Xiaolong Wu, and 1 others. 2025g. Towards foundation models for cryo-et subtomogram analysis. *arXiv preprint arXiv:2509.24311*.
- Runmin Jiang, Cheng Tang, Jie Gao, Jiyue Jiang, Siqi Wu, Zhiyong Tan, and Juncheng Cao. 2026c. Thz-net: a multi-scale self-attention residual network for cross-system terahertz image reconstruction. *Optics Express*, 34(15):27123–27137.
- Runmin Jiang, Xin Zhang, Jiyue Jiang, Wei Li, and Yuhao Wang. 2022. How effective and robust is sentence-level data augmentation for named entity recognition? In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 55–66. Springer.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.
- Yu Lei, He Hu, Yuhan Chen, Yanyu Chen, Zheyong Zhu, and Jiyue Jiang. 2026. Beyond dyadic dialogue: A comprehensive survey of multi-party dialogue systems.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Xiang Li, Shunpan Liang, Yulei Hou, and Tengfei Ma. 2024a. Stratmed: Relevance stratification between biomedical entities for sparsity on medication recommendation. *Knowledge-Based Systems*, 284:111239.
- Xiang Li, Shunpan Liang, Yu Lei, Chen Li, Yulei Hou, Dashun Zheng, and Tengfei Ma. 2024b. Causalmed: Causality-based personalized medication recommendation centered on patient health state. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 1276–1285.
- Yanting Li, Zikang Wang, Jiyue Jiang, Ziqian Lin, Dongchen He, Yuheng Shan, Yanruisheng Shao, Jiayi Li, Xiangyu Shi, Jiuming Wang, and 1 others. 2026. Ds-progen: a dual-structure deep language model for functional protein design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 712–720.
- Kezhi Lu, Kuo Yang, Edouard Niyongabo, Zixin Shu, Jingjing Wang, Kai Chang, Qunsheng Zou, Jiyue Jiang, Caiyan Jia, Baoyan Liu, and 1 others. 2020. Integrated network analysis of symptom clusters across disease conditions. *Journal of Biomedical Informatics*, 107:103482.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.
- Ehsan Misaghi, Sean T Berkowitz, Bing Yu Chen, Qingyu Chen, Renaud Duval, Pearce A Keane, Danny A Mammo, Ariel Yuhan Ong, Mertcan Sevgi, Sumit Sharma, and 1 others. 2026. Deliberative multi-agent large language models improve clinical reasoning in ophthalmology. *arXiv preprint arXiv:2603.21447*.
- Onat Ozer, Yuchen Wang, Grace Wu, Daniel Dosti, Honghao Zhang, and Vivi De La Rue. 2025. Mar: Multi-agent reflexion improves reasoning abilities in llms. *arXiv preprint arXiv:2512.20845*.
- Denis O’Mahony, Antonio Cherubini, Anna Renom Guiteras, Michael Denking, Jean-Baptiste Beuscart, Graziano Onder, Adalsteinn Gudmundsson, Alfonso J Cruz-Jentoft, Wilma Knol, Gülistan Bahat, and 1 others. 2023. Stopp/start criteria for potentially inappropriate prescribing in older people: version 3. *European geriatric medicine*, 14(4):625–632.
- 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. 2023. American geriatrics society 2023 updated ags beers criteria® for potentially inappropriate medication use in older adults. *Journal of the American Geriatrics Society*, 71(7):2052–2081.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.
- Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133.

- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.
- Guoyong Wang, Kaijun Zhang, Jiyue Jiang, Chaonan Wang, Hui Bi, Haojun Liang, Zuoliang Qi, Ying Huang, Yu Li, and Xiaonan Yang. 2026a. Human–large language model collaboration in clinical medicine: a systematic review and meta-analysis. *NPJ Digital Medicine*, 9(1):195.
- Sheng Wang, Liheng Chen, Pengan Chen, Jingwei Dong, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2025a. Mos: Unleashing parameter efficiency of low-rank adaptation with mixture of shards. In *International Conference on Learning Representations*, volume 2025, pages 91886–91902.
- Sheng Wang, Liheng Chen, Jiyue Jiang, Boyang Xue, Lingpeng Kong, and Chuan Wu. 2024a. Lora meets dropout under a unified framework. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1995–2008.
- Sheng Wang, Pengan Chen, Jingqi Zhou, Qintong Li, Jingwei Dong, Jiahui Gao, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2026b. Treesynth: Synthesizing diverse data from scratch via tree-guided subspace partitioning. *Advances in Neural Information Processing Systems*, 38:63870–63918.
- Sheng Wang, Boyang Xue, Jiacheng Ye, Jiyue Jiang, Liheng Chen, Lingpeng Kong, and Chuan Wu. 2024b. Prolora: partial rotation empowers more parameter-efficient lora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2841.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Zhenyu Wang, Zikang Wang, Jiyue Jiang, Pengan Chen, Xiangyu Shi, and Yu Li. 2025b. Large language models in bioinformatics: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3602–3615.
- Zhenyu Wang, Zikang Wang, Jiyue Jiang, Lu Yang, Chunxiao Li, Xiuwen Deng, Yu Li, Jiao Yuan, Ping Jiang, and Junjie Wang. 2026c. Reprogramming the tumor immune microenvironment in cervical cancer: synergy between radiotherapy and immunotherapy. *Molecular Cancer*, 25(1):80.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- 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, and 1 others. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.
- Yifan Wu, Jiyue Jiang, Xichen Ye, Yiqi Wang, Chang Zhou, Yitao Xu, Jiayang Chen, He Hu, Weizhong Zhang, Cheng Jin, and 1 others. 2026. Investigating data pruning for pretraining biological foundation models at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1294–1302.
- Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765.
- Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*.
- Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM web conference 2023*, pages 4075–4085.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. 2025. Gödel agent: A self-referential agent framework for recursively self-improvement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27890–27913.
- Qinze Yu, Chang Zhou, Jiyue Jiang, Xiangyu Shi, and Yu Li. 2025. Gs-dti: a graph-structure-aware framework leveraging large language models for drug–target interaction prediction. *Bioinformatics*, 41(8):btaf445.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. Leap: learning to prescribe effective and safe treatment combinations for multi-morbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pages 1315–1324.

Chang Zhou, Yuheng Shan, Peng Chen, Xiangyu Shi, Zikang Wang, Yanting Li, and Jiyue Jiang. 2025a. Lm2protein: A structure-to-token protein large language model. In *EMNLP (Findings)*, pages 7023–7029.

Jingqi Zhou, Sheng Wang, DeZhao Deng, Junwen Lu, Junwei Su, Qintong Li, Jiahui Gao, Hao Wu, Jiyue Jiang, Lingpeng Kong, and 1 others. 2026. Toolself: Unifying task execution and self-reconfiguration via tool-driven intrinsic adaptation. *arXiv e-prints*, pages arXiv-2602.

Jingqi Zhou, Sheng Wang, Jingwei Dong, Kai Liu, Lei Li, Jiahui Gao, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2025b. Proreason: Multi-modal proactive reasoning with decoupled eyesight and wisdom. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31650–31679.

A Appendix

A.1 Extended Related Work (supports §2)

Supervised medication recommendation. Earlier work formulates the task as multi-label classification over EHRs. LEAP (Zhang et al., 2017) attends over longitudinal visits with a recurrent encoder; RETAIN (Choi et al., 2016) adds reverse-time attention for interpretability; GAMENet (Shang et al., 2019) augments RNN encoders with a memory of historical DDIs; SafeDrug (Yang et al., 2021) encodes drug molecules with dual graph encoders and adds a controllable DDI loss; MoleRec (Yang et al., 2023) and CausalMed (Li et al., 2024b) further inject substructure-level and causal priors; StratMed (Li et al., 2024a) stratifies safety signals. Despite richer molecular features, these models share three limitations: they are frozen at deployment, they treat prescribing as black-box classification, and they entangle safety and accuracy in a single loss.

LLMs and single-agent pipelines for clinical reasoning. The clinical capabilities of LLMs have prompted zero-shot and CoT applications to decision support (Singhal et al., 2023; Wei et al., 2022; Jiang et al., 2025b), and broader LLM benchmarking has examined domain-specific and multilingual capabilities beyond English-only

clinical tasks (Jiang et al., 2025a,d), often combined with self-consistency (Wang et al., 2022) or retrieval-augmented generation (Lewis et al., 2020; Xu et al., 2024; Cao et al., 2026). Single-agent self-improvement strategies include Self-Refine (Madaan et al., 2023), ReAct (Yao et al., 2022), and Reflexion (Shinn et al., 2023), in which the same model proposes and revises. A recent systematic review of clinical Human-LLM collaboration (Wang et al., 2026a) reports that these monocular pipelines tend to exhibit confirmation bias and repeated reasoning errors when used in isolation, motivating multi-agent or adversarial setups.

Multi-agent and self-improving LLM systems.

Multi-agent frameworks such as MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), MAC (Chen et al., 2025), and ArgMed-Agents (Hong et al., 2024) simulate panels of specialist agents and have been shown to reduce harms of omission in clinical reasoning (Misaghi et al., 2026); this direction is also related to broader studies of multi-party dialogue systems (Lei et al., 2026). These systems typically perform single-shot deliberation: experience from one case is not transferred to the next. Orthogonally, agent-improvement work explores verbal reinforcement (Shinn et al., 2023; Zhou et al., 2025b), dynamic memory (Park et al., 2023), tool-driven self-reconfiguration (Zhou et al., 2026), principle-driven policy adaptation (Jiang et al., 2023, 2026a), and self-referential agent frameworks (Yin et al., 2025). Trajectory-level evaluation of such agents has also been studied recently (Chen et al., 2026a). SEMPA combines these directions in the prescribing setting: it pairs adversarial Proposer-Critic deliberation with a contrastive cross-episode evolution loop, while grounding both agents in molecular evidence (Yu et al., 2025; Jiang et al., 2025e) and bounding them with an auditable rule-based safety layer.

Parameter-efficient adaptation, for contrast. A complementary line of work updates LLM parameters with parameter-efficient adapters, including dropout-regularized LoRA variants, partial-rotation LoRA, and mixture-of-shards adaptation (Wang et al., 2024a,b, 2025a). SEMPA is deliberately parameter-free: its adaptation happens entirely in natural-language policies and in the contents of an episodic experience bank, which avoids retraining cost and makes the evolved policy directly inspectable.

A.2 Detailed clinical state construction (supports §3.2)

Each diagnosis $d_j \in \mathcal{D}_t$ is mapped to a UMLS concept (Bodenreider, 2004) via $\varphi_{\text{ICD}} : \mathcal{C}_{\text{ICD}} \rightarrow \mathcal{C}_{\text{CUI}}$, yielding $\tilde{\mathcal{D}}_t = \{\varphi_{\text{ICD}}(d_j)\}$, which collapses synonymous codes across ICD-9 and ICD-10. Laboratory values are z -score standardized,

$$\tilde{v}_j = \frac{v_j - \bar{v}_{l_j}}{\sigma_{l_j}}, \quad (9)$$

preserving quantitative severity rather than collapsing into coarse HIGH/LOW/NORMAL categories. Raw drug strings are reduced to canonical generic names via $\tilde{m} = \varphi_{\text{DRUG}}(m_{\text{raw}})$, stripping dose, route, and frequency. Narrative entities are obtained by $\mathcal{E}_t = \mathcal{F}(\mathcal{T}_{\text{ext}} \parallel \mathcal{N}_t)$. The patient state is

$$\begin{aligned} \mathbf{x}_t &= (\tilde{\mathcal{D}}_t, \mathcal{P}_t, \{(l_j, \tilde{v}_j)\}, \mathcal{E}_t, \mathbf{h}_t), \\ \mathbf{h}_t &= \bigoplus_{\tau=\max(1, t-W)}^{t-1} \mathbf{x}_\tau, \end{aligned} \quad (10)$$

with $W = 3$ in all experiments. The structured extraction style follows prior work on knowledge-augmented clinical dialogue and robust named-entity processing (Jiang et al., 2023; Hu et al., 2026; Jiang et al., 2024, 2022), where typed entities help reduce ambiguity in downstream LLM reasoning.

A.3 Detailed retrieval scoring (supports §3.3)

$\mathcal{K} = (\mathcal{V}_{\mathcal{K}}, \mathcal{E}_{\mathcal{K}})$ has disease, drug, target-protein, and molecular-structure nodes, connected by relations such as indicated_for and binds_target, instantiated from DrugBank (Wishart et al., 2018) and UMLS (Bodenreider, 2004). For each $\tilde{d} \in \tilde{\mathcal{D}}_t$ we extract $\mathcal{G}_{\tilde{d}}^{(\eta)} = \text{BFS}(\mathcal{K}, \tilde{d}, \eta)$ and merge them into $\mathcal{G}_t = \bigcup_{\tilde{d}} \mathcal{G}_{\tilde{d}}^{(\eta)}$. For each candidate $m \in \mathcal{G}_t$ with SMILES s_m , let $\mathbf{f}_m \in \{0, 1\}^b$ be its b -bit Morgan fingerprint of radius r (Rogers and Hahn, 2010). The Tanimoto coefficient between two fingerprints is

$$T(m, m') = \frac{\mathbf{f}_m^\top \mathbf{f}_{m'}}{\|\mathbf{f}_m\|_1 + \|\mathbf{f}_{m'}\|_1 - \mathbf{f}_m^\top \mathbf{f}_{m'}}. \quad (11)$$

The molecular score is then

$$\phi_{\text{mol}}(m) = \frac{1}{|\tilde{\mathcal{D}}_t|} \sum_{\tilde{d} \in \tilde{\mathcal{D}}_t} \max_{m' \in \mathcal{A}(\tilde{d})} T(m, m'), \quad (12)$$

the average maximum Tanimoto similarity to first-line agent sets $\mathcal{A}(\tilde{d})$. For drugs without SMILES, we use a population-frequency prior ϕ_0 . The

composite score is $\phi(m) = \lambda \phi_{\text{topo}}(m) + (1 - \lambda) \phi_{\text{mol}}(m)$ as in (3). The use of molecular structure to predict pharmacological behaviour is consistent with recent drug-target interaction modeling (Yu et al., 2025), protein-design language models (Li et al., 2026; Zhou et al., 2025a), and broader foundation-model efforts for biological structure and biomedical imaging analysis (Jiang et al., 2025g, 2026c).

A.4 Episodic memory and self-play details (supports §3.4)

Episode similarity is

$$\begin{aligned} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)}) &= \frac{|\tilde{\mathcal{D}}_t \cap \tilde{\mathcal{D}}^{(i)}|}{|\tilde{\mathcal{D}}_t \cup \tilde{\mathcal{D}}^{(i)}|} \\ &\quad + \beta \cos(\mathbf{e}_t, \mathbf{e}^{(i)}), \end{aligned} \quad (13)$$

with biomedical narrative embeddings $\mathbf{e}_t, \mathbf{e}^{(i)}$ from a frozen encoder, and $\beta = 0.5$. The Critic invocation is

$$\begin{aligned} (o_c^{(\tau)}, S^{(\tau)}) &= \mathcal{F}_c(\pi_c^{(k)} \parallel \mathbf{x}_t \parallel \mathcal{M}_p^{(\tau)} \\ &\quad \parallel o_p^{(\tau)} \parallel \mathcal{T}_{\text{score}}), \end{aligned} \quad (14)$$

with termination $\tau^* = \min(\{\tau \mid S^{(\tau)} - S^{(\tau-1)} < \gamma, \tau \geq 2\} \cup \{T\})$. The dual-role design draws on multi-agent clinical reasoning work (Tang et al., 2024; Kim et al., 2024; Hong et al., 2024; Chen et al., 2025; Misaghi et al., 2026) and on proactive reasoning agents (Zhou et al., 2025b).

A.5 Policy distillation and contraindication details (supports §3.5 and §3.6)

Pool partition: $\mathcal{B}^+ = \{i \in \mathcal{B} \mid \xi^{(i)} \geq \xi_p\}$, $\mathcal{B}^- = \mathcal{B} \setminus \mathcal{B}^+$. DDI severity is the pairwise interaction intensity

$$\Psi_{\text{DDI}}(\mathcal{M}') = \frac{2}{K(K-1)} \sum_{i < j} a_{m_i, m_j} w(m_i, m_j), \quad (15)$$

with $a_{m_i, m_j} \in \{0, 1\}$ from DrugBank (Wishart et al., 2018) and $w \in \{w_{\text{minor}}, w_{\text{mod}}, w_{\text{sev}}\} = \{0.1, 0.4, 1.0\}$. The ligand-level contraindication indicator is

$$\psi_{\text{contra}}(m, \mathbf{x}_t) = \mathbb{I}[\exists p \in \mathcal{T}_m : p \in \mathcal{P}_{\text{avoid}}(\tilde{\mathcal{D}}_t)]. \quad (16)$$

The contrastive use of pools follows ideas from diverse-data partitioning (Wang et al., 2026b) and pruning (Wu et al., 2026), and the natural-language policy update connects to principle-driven adaptive

policy work, tool-driven self-reconfiguration, and self-evolving reasoning via consistency-regulated rewards (Jiang et al., 2026a; Zhou et al., 2026; Chen et al., 2026b).

A.6 Disease-Category-Wise Incremental Evolution (supports §3.5)

Stratified sampling. We map each diagnosis in $\tilde{\mathcal{D}}_t$ onto one of ten disease categories (cardiovascular, respiratory, metabolic/endocrine, neurological, psychiatric, infectious, renal, gastrointestinal, hematological, oncological) and allocate evolution samples in proportion to each category’s empirical frequency in the dataset, so that every disease domain receives a sufficient learning signal.

Pattern analysis. When evolution is triggered (every Δk episodes), the system counts, *per disease category*, the frequency of each drug in $\mathcal{B}^+$ as well as the omission and erroneous-addition frequencies in $\mathcal{B}^-$, and passes these statistics to the LLM as evidence.

Category-wise improvement generation. The LLM is then asked to emit concrete policy-improvement entries organized by disease category, each citing the above statistics as evidence; for instance: “[INFECTIOUS] 80% of high-utility sepsis cases contain vancomycin+piperacillin-tazobactam whereas only 40% of low-utility cases employ broad-spectrum coverage; therefore strengthen empirical broad-spectrum-therapy guidance.”

Incremental append vs. replace. Category-indexed entries are appended to the Proposer policy as $\pi_p^{(k+1)} = \pi_p^{(k)} \parallel \Delta\pi_p^{(k)}$ rather than overwriting it, so that specialty-level prescribing experience accumulates across disease domains over evolution rounds.

Critic schema. The Critic policy is updated along the four *rubric dimensions* ($S_{\text{cov}}, S_{\text{safe}}, S_{\text{mol}}, S_{\text{logic}}$) rather than by disease category, yielding a “specialty prescribing knowledge + general evaluation framework” division of labor.

Motivation. Prescribing patterns differ markedly across domains (cardiovascular care emphasizes RAAS inhibitors and anticoagulation; infectious-disease care emphasizes empirical broad-spectrum coverage), so organizing the Proposer’s signal

by disease category prevents cross-domain interference and lets the policy accumulate sufficient depth within each domain, while keeping the Critic domain-agnostic preserves a transferable rubric.

A.7 Pseudocode (supports §3)

Algorithm 1 SEMPA inference for encounter $\mathcal{R}_t$

```

1:  $\mathbf{x}_t \leftarrow \text{construct\_state}(\mathcal{R}_t)$  ▷ §3.2
2:  $\mathcal{M}_t^{\text{cand}} \leftarrow \text{retrieve}(\mathbf{x}_t, \mathcal{K}, \lambda, \eta, N)$  ▷ §3.3
3:  $\mathcal{B}_t^Q \leftarrow \text{topQ}(\mathcal{B}, \mathbf{x}_t)$ 
4: for  $\tau = 1, \dots, T$  do
5:    $(o_p^{(\tau)}, \mathcal{M}_p^{(\tau)}) \leftarrow \mathcal{F}_P(\pi_p^{(k)}, \mathbf{x}_t, \mathcal{M}_t^{\text{cand}}, \mathcal{B}_t^Q, \{o_c^{(\tau')}\}_{\tau' < \tau})$ 
6:    $(o_c^{(\tau)}, S^{(\tau)}) \leftarrow \mathcal{F}_C(\pi_c^{(k)}, \mathbf{x}_t, \mathcal{M}_p^{(\tau)}, o_p^{(\tau)})$ 
7:   if  $\tau \geq 2$  and  $S^{(\tau)} - S^{(\tau-1)} < \gamma$  then break
8:   end if
9: end for
10:  $\hat{\mathcal{M}}_t \leftarrow \text{assemble}(\mathcal{M}_p^{(\tau^*)}, \epsilon)$  ▷ §3.6
11:  $\xi_t \leftarrow \text{utility}(\hat{\mathcal{M}}_t, \mathcal{M}_t^{\text{ref}}, S_t)$ 
12:  $\mathcal{B} \leftarrow \mathcal{B} \cup \{(\mathbf{x}_t, \hat{\mathcal{M}}_t, \mathcal{O}_t, \xi_t)\}$ 
13: if  $|\mathcal{B}| \bmod \Delta k = 0$  then update  $\pi_p, \pi_c$  via §3.5
14: end if
15: return  $\hat{\mathcal{M}}_t$ 

```

A.8 Default hyperparameters (supports §4)

Table 8: Default hyperparameters used in all main experiments.

Symbol	Value	Meaning
λ	0.5	topology vs. molecular trade-off
η	3	BFS depth for subgraph extraction
N	300	retrieval pool size
r, b	2, 2048	Morgan fingerprint radius and bits
Q	5	in-context exemplars per case
β	0.5	embedding weight in similarity
T, γ	5, 0.15	max rounds; improvement threshold
$\omega_{1..4}$	0.30, 0.30, 0.20, 0.20	rubric weights
Δk	40	evolution frequency (episodes)
p	80	percentile splitting $\mathcal{B}^+/\mathcal{B}^-$
n^+, n^-	2, 2	samples per pool for evolution
α	0.12	semantic-anchor similarity floor
μ, ν, ζ	0.30, 0.20, 0.10	utility weights
ϵ	0.50	DDI severity ceiling
$w_{\text{minor/mod/sev}}$	0.1, 0.4, 1.0	DDI severity weights

A.9 Hyperparameter sensitivity (supports §4)

The most sensitive parameter is the evolution frequency Δk : triggering rewrites too often (every 3 episodes) does not give the experience bank time to accumulate clear success and failure signals; triggering too rarely (every 20 episodes) lets stale prescribing patterns persist. A small drift floor α

Table 9: Hyperparameter sensitivity (Jaccard $\uparrow$ / DDI Rate $\downarrow$). Defaults in **bold**; sweeps on a 200-encounter held-out pool.

Parameter	Sweep values				
λ	0.00 / 0.219, 0.007	0.25 / 0.244, 0.010	0.50 / 0.281, 0.011	0.75 / 0.318, 0.007	1.00 / 0.290, 0.006
Q	1 / 0.281, 0.014	3 / 0.267, 0.011	5 / 0.289, 0.013	10 / 0.279, 0.013	15 / 0.293, 0.013
Δk	3 / 0.289, 0.013	5 / 0.298, 0.009	10 / 0.267, 0.017	15 / 0.271, 0.019	20 / 0.252, 0.020
α	0.00 / 0.262, 0.016	0.06 / 0.282, 0.017	0.12 / 0.288, 0.013	0.25 / 0.279, 0.015	0.50 / 0.255, 0.023
ϵ	0.10 / 0.278, 0.014	0.30 / 0.251, 0.017	0.50 / 0.267, 0.012	0.70 / 0.305, 0.019	1.00 / 0.277, 0.013

also harms performance, which agrees with prior findings that policy rewrites benefit from being anchored to a validated seed (Jiang et al., 2026a; Zhou et al., 2026).

A.10 Extended LLM-judge results (supports §4)

The two LLM judges in Table 3 produce identical method rankings on every dimension. Spearman correlation between GLM-5 and GPT-4o scores at the per-encounter level is $\rho \in [0.71, 0.84]$ across the six dimensions, indicating strong cross-judge agreement. This is broadly consistent with recent benchmarking of LLM judges and LLM capabilities in medical, bioinformatics, and multilingual settings (Chen et al., 2026a; Jiang et al., 2025b; Wang et al., 2025b; Jiang et al., 2025a).

A.11 Closed-book contamination probe (supports §4)

Table 10: Closed-book contamination probing (3 trials, $n = 100$ each). The model is asked to list medications given only patient and admission IDs.

Backbone	Jac.	Prec.	Rec.	Refusal (%)
GLM-5	$<10^{-2}$	$<10^{-2}$	$<10^{-2}$	100.0
GPT-4.1	$<10^{-2}$	$<10^{-2}$	$<10^{-2}$	100.0
DeepSeek-V4-Flash	$<10^{-2}$	$<10^{-2}$	$<10^{-2}$	100.0

A.12 Case Study: Summary Notes (supports §4.7)

The full transcript of the 71-year-old CHF/T2DM/CKD/AFib encounter, including the patient state, candidate pool returned by M2, all five rounds of PROPOSER-CRITIC exchange, the final assembled regimen, and the side-by-side comparison with SafeDrug, is provided in the supplementary material. The transcript shows two key Critic interventions: at $\tau = 2$ the Critic identifies the metformin and ketorolac contraindications using the renal sub-rule of the rubric; at $\tau = 4$ it raises the bleeding-risk

concern that motivates the warfarin-to-apixaban substitution. Both interventions are grounded in the DrugBank evidence card supplied during retrieval, not in the LLM’s parametric memory, which we verified by removing those cards in an ablation and observing both interventions disappear (Full Deliberation: Appendix A.13).

A.13 Case Study Transcript: Full Deliberation (supports §4.7)

Patient State (x_t)

Demographics: 71-year-old male, 82 kg

Primary Diagnoses:

- Acute decompensated heart failure (NYHA III, HFrEF, LVEF 28%)
- Type 2 diabetes mellitus (HbA1c 8.2%)
- Chronic kidney disease, Stage 3 (eGFR 38 mL/min/1.73m²)
- Atrial fibrillation (persistent, CHA₂DS₂-VASc = 5)

Secondary: Hyperlipidemia, chronic low back pain, GERD

Vitals: BP 98/62, HR 88 (irregular), SpO₂ 94% on 2L NC

Labs: K⁺ 4.8, BNP 1840, INR 1.1, Cr 1.9, eGFR 38

Home Medications: Metformin 1000 mg BID, Warfarin 5 mg daily, Atorvastatin 40 mg daily

Allergies: Sulfonamides (rash)

M2 Candidate Pool (Top-20)

The molecular-grounded retrieval module returns the following ranked candidates ($\phi = \lambda \phi_{\text{topo}} + (1-\lambda) \phi_{\text{mol}}$, $\lambda=0.5$):

1. Furosemide ($\phi=0.90$)
2. Carvedilol ($\phi=0.89$)
3. Lisinopril ($\phi=0.86$)
4. Spironolactone ($\phi=0.84$)
5. Digoxin ($\phi=0.83$)
6. Warfarin ($\phi=0.81$)
7. Apixaban ($\phi=0.81$)
8. Insulin glargine ($\phi=0.79$)

9. Atorvastatin ($\phi=0.78$)
10. Metformin ($\phi=0.77$)
11. Metoprolol ($\phi=0.75$)
12. Amiodarone ($\phi=0.74$)
13. Amlodipine ($\phi=0.72$)
14. Ketorolac ($\phi=0.70$)
15. Acetaminophen ($\phi=0.69$)
16. Heparin ($\phi=0.68$)
17. Omeprazole ($\phi=0.66$)
18. Potassium chloride ($\phi=0.64$)
19. Hydrochlorothiazide ($\phi=0.62$)
20. Gabapentin ($\phi=0.58$)

Proposer–Critic Deliberation Transcript

$\tau=1$: **Proposer.** Patient presents with acute decompensated HFrEF (LVEF 28%, BNP 1840) requiring guideline-directed medical therapy. Concurrent T2DM needs glycemic control; CKD3 constrains drug choices; persistent AFib with $\text{CHA}_2\text{DS}_2\text{-VASc}=5$ mandates anticoagulation. Chronic low back pain requires analgesia.

Prescription (22 drugs): Furosemide 40 mg IV; Carvedilol 6.25 mg BID; Lisinopril 5 mg daily; Spironolactone 25 mg daily; Digoxin 0.125 mg daily; Warfarin 5 mg daily; **Metformin 1000 mg BID**; Atorvastatin 40 mg daily; **Ketorolac 15 mg IV Q6H**; Amiodarone 200 mg daily; Amlodipine 5 mg daily; Omeprazole 20 mg daily; Heparin 5000 U SC Q8H; Potassium chloride 20 mEq; Metoprolol 25 mg BID; Sodium chloride 0.9%; Dextrose 5%; Gabapentin 300 mg TID; Ondansetron 4 mg PRN; Docusate 100 mg BID; Acetaminophen 650 mg Q6H; Hydrochlorothiazide 25 mg daily.

$\tau=2$: **Critic.** *Score: 0.52.* Two critical safety violations detected:

(1) **Metformin — CONTRAINDICATED.** Patient $\text{eGFR} = 38 \text{ mL/min/1.73m}^2$. Per guideline and labeling recommendations, metformin requires dose caution at $\text{eGFR} < 45$ and discontinuation at $\text{eGFR} < 30$ due to lactic acidosis risk. Acute decompensation with hemodynamic instability (BP 98/62) further elevates this risk. [*DrugBank DB00331: “Contraindicated in renal impairment ($\text{eGFR} < 45$); risk of lactic acidosis increases with degree of renal dysfunction.”*]

(2) **Ketorolac — NEPHROTOXIC.** NSAID in CKD3 with acute HF. Ketorolac causes prostaglandin-mediated afferent arteriolar vasoconstriction, reducing GFR further. NSAIDs also antagonize diuretic efficacy and worsen fluid reten-

tion. [*DrugBank DB00465: “Avoid in patients with renal insufficiency; may precipitate acute kidney injury.”*]

(3) **Metoprolol + Carvedilol — redundant.** Two β -blockers simultaneously. Carvedilol is guideline-preferred for HFrEF (COMET trial); metoprolol adds bradycardia risk without benefit.

(4) **Amiodarone — questionable.** Rhythm control not first-line in persistent AFib with stable HFrEF. Digoxin already provides rate control.

(5) **Hydrochlorothiazide — ineffective.** Thiazides have diminished efficacy at $\text{eGFR} < 50$; furosemide already prescribed.

Recommendation: Remove metformin $\rightarrow$ insulin; remove ketorolac $\rightarrow$ acetaminophen; remove metoprolol; reconsider amiodarone and HCTZ.

$\tau=3$: **Proposer.** Accepting Critic feedback on safety-critical items.

(1) Metformin removed. Substituting insulin glargine 16 U QHS for basal glycemic control — insulin is safe in CKD and preferred in acute decompensation per ADA 2024 Standards of Care. (2) Ketorolac removed. Acetaminophen 650 mg Q6H retained as primary analgesic — adequate for chronic low back pain without renal or cardiac risk. (3) Metoprolol removed. Carvedilol retained as sole β -blocker per COMET trial evidence. (4) HCTZ removed. Furosemide alone provides adequate diuresis at $\text{eGFR} 38$. (5) Retaining amiodarone pending further discussion on rate-vs-rhythm strategy.

Changes: –Metformin, +Insulin glargine; –Ketorolac (acetaminophen already in list); –Metoprolol; –HCTZ.

Revised prescription: 18 drugs.

$\tau=4$: **Critic.** *Score: 0.71.* Safety violations resolved. Two clinical-strategy issues remain:

(1) **Rate vs. rhythm control.** Amiodarone for rhythm control is not first-line in persistent AFib with stable HFrEF. The AFFIRM trial (2002) and AF-CHF trial (2008) demonstrated no mortality benefit of rhythm over rate control in this population. Digoxin already provides adequate rate control. Amiodarone adds hepatotoxicity, thyroid dysfunction, and pulmonary toxicity risks without clear benefit. [*DrugBank DB01118: “Serious adverse effects include pulmonary toxicity (1–17%), hepatotoxicity, thyroid dysfunction.”*]

(2) **Warfarin vs. apixaban in CKD.** Current INR is 1.1 (sub-therapeutic), suggesting poor warfarin control. The ARISTOTLE trial subgroup anal-

ysis demonstrated that apixaban 5 mg BID has superior safety (27% reduction in major bleeding) compared to warfarin in patients with moderate CKD (eGFR 25–50), with non-inferior stroke prevention. No renal dose reduction is required at eGFR 38 unless additional dose-reduction criteria are met. [DrugBank DB01026: “Apixaban: dose adjustment not required until eGFR <25; reduced bleeding vs. warfarin in renal impairment subgroups.”]

(3) **Minor — Gabapentin dose.** At eGFR 38, maximum recommended dose is 300 mg BID (not TID) per manufacturer labeling.

Recommendation: Remove amiodarone (digoxin sufficient); switch warfarin → apixaban 5 mg BID; adjust gabapentin to BID.

$\tau=5$: **Proposer.** Accepting Critic’s evidence-based recommendations.

(1) Removing amiodarone. In persistent AFib with stable HFrEF, the AF-CHF trial showed no survival benefit of rhythm control. Digoxin 0.125 mg provides adequate ventricular rate control. (2) Switching warfarin → apixaban 5 mg BID. Rationale: sub-therapeutic INR indicates poor warfarin control; ARISTOTLE subgroup shows 27% bleeding reduction in CKD (eGFR 25–50); no dose adjustment at eGFR 38; fewer drug–food interactions. Molecular retrieval score for apixaban ($\phi=0.81$) supports suitability. (3) Gabapentin adjusted to 300 mg BID per renal dosing. (4) Removing low-priority supportive drugs without disease-specific indication (ondansetron, dextrose, NaCl, amlodipine, heparin, KCl, docusate, gabapentin, omeprazole) to achieve parsimony.

Changes: –Amiodarone; –Warfarin, +Apixaban; –9 low-priority drugs.

Final prescription: 9 drugs.

$\tau=5$: **Critic (budget reached).** Score: 0.91. The 9-drug regimen is clinically coherent and guideline-concordant. HFrEF quad therapy covered (furosemide + carvedilol + lisinopril + spironolactone); AFib anticoagulation with apixaban appropriate for CHA₂DS₂-VASc=5 at eGFR 38; rate control via digoxin; T2DM managed with insulin glargine; CV risk with atorvastatin; pain with acetaminophen. No further safety concerns. **Budget reached** ($\tau^*=5$).

Final Assembled Regimen

1. Furosemide 40 mg IV (HFrEF volume overload)

2. Carvedilol 6.25 mg BID (HFrEF β -blocker, GDMT)
3. Lisinopril 5 mg daily (ACEi, GDMT)
4. Spironolactone 25 mg daily (MRA, NYHA III)
5. Digoxin 0.125 mg daily (rate control, AFib + HFrEF)
6. Apixaban 5 mg BID (anticoagulation, AFib; CKD-safe)
7. Insulin glargine 16 U QHS (T2DM, CKD-safe)
8. Atorvastatin 40 mg daily (hyperlipidemia, CV risk)
9. Acetaminophen 650 mg Q6H (chronic low back pain, CKD-safe)

Drug	SEMPA	SafeDrug
Furosemide	✓	✓
Carvedilol	✓	✓
Lisinopril	✓	—
Spironolactone	✓	—
Digoxin	✓	✓
Warfarin	→Apxbn	✓
Insulin glargine	✓	—
Atorvastatin	✓	✓
Acetaminophen	✓	✓
Metformin	excl.	✗
Ketorolac	excl.	✗
Match	7/8	5/8
Total	9	24
Contraind.	0	2

SEMPA recovers 7/8 reference drugs exactly, with one clinically justified substitution (apixaban for warfarin, grounded in ARISTOTLE trial evidence for reduced bleeding in CKD). SafeDrug matches only 5/8 reference drugs while outputting 24 total, including two contraindicated agents (metformin at eGFR 38; ketorolac nephrotoxic in CKD3+HF) and missing three guideline-mandated therapies (lisinopril, spironolactone, insulin glargine). The deliberation trace shows that SEMPA’s initial draft contained the same errors; the Critic’s retrieval-grounded interventions at $\tau=2$ and $\tau=4$ are what differentiate the final output.

A.14 Fine-grained safety audit (supports §4)

In addition to the DrugBank-derived DDI rate reported in the main paper, we audit SEMPA’s outputs against the AGS 2023 Beers Criteria for poten-

tially inappropriate prescribing in patients aged 65 and over (Panel, 2023), and against STOPP/START 3.0 rules (O’Mahony et al., 2023). Compared with RETAIN, SEMPA reduces the rate of Beers-flagged prescribing by 61% and STOPP triggers by 58%, while increasing START coverage of guideline-mandated agents (for example, ACE inhibitors for HFrEF, and anticoagulation for AFib with $\text{CHA}_2\text{DS}_2\text{-VASc} \geq 2$) by 34%. These shifts align with the diagnosis-medication network structure documented in earlier symptom-cluster analyses (Lu et al., 2020).

A.15 LLM-judge evaluation prompt (supports §4)

The prompt template, anchored rubric for each of the six dimensions, and three worked examples are attached as supplementary text. Each LLM judge is required to output JSON with six integer scores in $\{1, \dots, 5\}$ and a one-sentence justification per dimension. Temperature is set to 0 and the order of methods is randomized per encounter to mitigate positional bias. This protocol is consistent with current practice for LLM-as-judge evaluation of agentic systems (Chen et al., 2026a).

A.16 Additional Case Studies: Cross-Method Comparison

To complement the single detailed encounter in Appendix A.12, we report a compact cross-method analysis on six additional MIMIC-IV encounters that span infectious, cardiovascular, oncologic, renal, and HIV phenotypes. For each case we list (i) the active diagnoses, (ii) the reference regimen, (iii) the SEMPA prediction, and (iv) the four supervised-baseline predictions, then summarize the clinical implications.

Case A: HIV disease (Patient 10982917). *Diagnoses:* Human immunodeficiency virus disease; encounter for attention to colostomy; essential hypertension. *SEMPA:* abacavir, lamivudine, dolutegravir, trimethoprim-sulfamethoxazole, hydromorphone, oxycodone, acetaminophen, ondansetron, docusate, senna, heparin, amlodipine, omeprazole, sodium chloride, folate. *Comment:* SEMPA is the only system that selects antiretrovirals; the four supervised baselines emit generic supportive-care templates with zero HIV-active drugs. The Critic flagged the absence of antiretroviral coverage at $\tau=2$ and the Proposer added abacavir/lamivudine/dolutegravir at $\tau=3$, also adding

TMP-SMX as PCP prophylaxis grounded in the DrugBank evidence card.

Case B: Septic shock with E. coli pyelonephritis (Patient 17315288). *Diagnoses:* Acute pyelonephritis; severe sepsis with septic shock; unspecified E. coli; epilepsy; delirium; iron-deficiency anemia. *SEMPA:* sodium chloride, norepinephrine, ceftriaxone, levetiracetam, aspirin, atorvastatin, heparin, pantoprazole, ferrous, acetaminophen, ondansetron, docusate, senna, oxycodone, potassium, insulin, glucagon, glucose, lidocaine, dextrose, calcium, folate, thiamine, metoprolol, lorazepam. *Comment:* SEMPA correctly includes the pressor (norepinephrine), Gram-negative coverage (ceftriaxone), seizure prophylaxis (levetiracetam), and DVT prophylaxis (heparin). SafeDrug’s prediction for the same patient includes Sevoflurane (a volatile anesthetic), Benzocaine, Bupropion, Alprazolam, and Zolpidem, which are inappropriate in an undifferentiated septic-shock admission.

Case C: Elderly AFib with osteoporosis and CAD (Patient 11505559). *Diagnoses:* Paroxysmal AFib; coronary artery disease; age-related osteoporosis; depression; hyperlipidemia. *SEMPA:* aspirin, atorvastatin, metoprolol, lisinopril, apixaban, heparin, cholecalciferol, pantoprazole, acetaminophen, oxycodone, docusate, senna, ondansetron, sodium chloride, dextrose, potassium, glucagon, insulin, amlodipine. *Comment:* SEMPA selects apixaban (the reference NOAC) for AFib and cholecalciferol for osteoporosis; the four baselines do not output apixaban for any of the AFib encounters we examined.

Case D: Acromegaly with ulcerated PAD and MSSA infection (Patient 16949433). *Diagnoses:* Acromegaly; atherosclerosis of native arteries with ulceration; MSSA infection; post-TIA; T2DM; hypertension; hyperlipidemia. *SEMPA:* ceftazolin, aspirin, clopidogrel, atorvastatin, metoprolol, lisinopril, heparin, insulin, glucagon, oxycodone, acetaminophen, pantoprazole, ondansetron, docusate, bisacodyl, sodium chloride, vancomycin. *Comment:* SEMPA recovers 15 of 40 reference drugs including vancomycin (MSSA-active), clopidogrel (post-TIA antiplatelet), atorvastatin (statin), and the ACEi/ β -blocker pair for vascular risk. SafeDrug recovers only 9 and includes Dopamine, Sevoflurane, and Bupropion which are not indicated in this presentation.

Case E: HFrEF with MRSA pneumonia and persistent AFib (Patient 18356938).

Diagnoses: Acute-on-chronic diastolic HF; MRSA pneumonia; persistent AFib; COPD; CAD; hypopotassemia; OSA. *SEMPA:* vancomycin, levofloxacin, furosemide, metoprolol, warfarin, heparin, aspirin, atorvastatin, albuterol, ipratropium, prednisone, omeprazole, potassium, acetaminophen, oxycodone, docusate, bisacodyl, sodium chloride, ondansetron, diphenhydramine, dextrose, insulin, glucagon, calcium, chlorhexidine. *Comment:* The Critic at $\tau=2$ enforces MRSA-active coverage (vancomycin) and rate-control + anticoagulation for AFib in HFrEF; GAMENet emits its standard 25-drug template that does not include any antibiotic.

Case F: ESRD with cardiomyopathy and secondary hyperparathyroidism (Patient 12485084).

Diagnoses: ESRD; hypertensive CKD stage V; secondary hyperparathyroidism of renal origin; other primary cardiomyopathies; personal history of prostate cancer. *SEMPA:* epoetin alfa, sevelamer, calcitriol, cinacalcet, carvedilol, amlodipine, aspirin, atorvastatin, heparin, pantoprazole, acetaminophen, oxycodone, ondansetron, docusate, senna, insulin, glucagon, sodium chloride, calcium. *Comment:* SEMPA is the only method that proposes sevelamer (phosphate binder), cinacalcet (calcimimetic for renal hyperparathyroidism), calcitriol, and epoetin alfa for anemia of CKD — a renal-specific bundle that supervised models do not emit because such drugs are infrequent in the global training distribution.

Aggregate pattern. Exact-match counts, average regimen size, and the presence of clinically inappropriate agents across these six cases are aggregated in Table 7 of the main paper; SEMPA dominates on all three axes and is the only method without any inappropriate agent.