

SafeRx-Agent: A Knowledge-Grounded Multi-Agent Framework for Safe and Explainable Medication Recommendation

Xinyu Wang^{2*}, Hanwei Wu^{1,3*}, Zhenghan Tai^{1,4*}, Sicheng Lyu², Qincheng Lu²,
Ziyu Zhao², Jijun Chi⁴, Jingrui Tian², Xiao-Wen Chang², Ziyang Song^{5†}

¹Humani Labs ²McGill University ³McMaster University

⁴University of Toronto ⁵Ohio University

Correspondence: ziyangs@ohio.edu

Abstract

Medication recommendation predicts medications for patient visits, but existing methods still face two key challenges. At the model level, traditional drug recommendation methods only predict structured drug codes with limited evidence grounding, while LLM agents can use richer clinical context but may lack safety verification and traceability. At the task level, existing benchmarks often use broad medication categories, which ignore subgroup-level safety differences and can lead to risk overestimation. We introduce the first fine-grained medication recommendation setting based on fourth-level ATC code generation. We propose **Safe Prescription Agent (SafeRx-Agent)**, a knowledge-grounded multi-agent framework that uses patient context, external clinical knowledge, and safety verification to recommend traceable medication sets. Experimental results on MIMIC-III and MIMIC-IV datasets show that SafeRx-Agent improves fine-grained medication prediction accuracy while controlling drug interactions, contraindications, and medication set size. Code accompanying this work is available at <https://github.com/HWu3150/SafeRx-Agent/tree/main>.

1 Introduction

Medication recommendation from electronic health records (EHRs) is a high-stakes clinical natural language processing (NLP) task (Xu et al., 2022). Based on a patient’s longitudinal clinical context, including prior visits, diagnoses, procedures, and medication history, the model predicts the medications for the current encounter (Song et al., 2026). This task is challenging because ICU histories are sparse, visits involve multiple active conditions, and prescribing depends on both acute illness and treatment continuity (Shang et al., 2019).

Deep learning methods have advanced medication prediction tasks by modeling patient visit sequences and drug co-occurrence patterns (Song et al., 2022; Ali et al., 2023), but they operate over structured EHR codes and cannot utilize patient-specific textual context or external medical evidence at inference time (Yang et al., 2021; Song et al., 2024). Large language models (LLMs) can process clinical text and generate medication sets, and multi-agent systems further decompose complex clinical reasoning into coordinated steps with tool use (Liu et al., 2025; Fan et al., 2025; Wang et al., 2026). However, unconstrained LLM agents remain unreliable for clinical decision support: prior work reports hallucinations, guideline misalignment, and unsafe recommendations (Hager et al., 2024; Asgari et al., 2025; Farrag et al., 2026). Safe medication recommendation therefore requires a knowledge-grounded agentic framework with explicit evidence and safety verification.

Medications are standardized with the Anatomical Therapeutic Chemical (ATC) taxonomy (WHO Collaborating Centre for Drug Statistics Methodology, 2026), which organizes drugs into five hierarchical levels. Most benchmarks predict at the third level (ATC-L3), i.e., the pharmacological subgroup level, which merges chemical subgroups that can differ in clinical use and safety profile (Ali et al., 2023). We instead predict at the finer-grained ATC-L4, the chemical subgroup level, which better reflects clinically relevant prescribing distinctions and at which safety relations are documented (Tatonetti et al., 2012; Kass-Hout et al., 2016). Collapsing a documented relation to its ATC-L3 parents propagates the flag to every sibling subgroup pair beneath them: B01AB (heparin group) $\times$ J07BB (influenza vaccines) carries no documented interaction at ATC-L4, yet the parent pair B01A $\times$ J07B is flagged because the interaction is recorded for a different child, J07BC (hepatitis vaccines). Across the cohort, collapsing ATC-L4

*Equal contribution.

†Corresponding author.

codes to ATC-L3 inflates the measured interaction rate of the same ground-truth prescriptions from 30.13% to 63.98% on MIMIC-IV (Section 6.7), so accurate safety measurement requires ATC-L4 granularity.

We propose **Safe Prescription Agent (SafeRx-Agent)**, a knowledge-grounded multi-agent framework for safe and explainable fine-grained medication recommendation. SafeRx-Agent routes each patient case to specialty-aware medication agents that generate fine-grained ATC-L4 medication candidates grounded in patient context, ICD and ATC taxonomies, and medication indication evidence. A safety-aware critic-revision loop then checks candidates against drug-drug interaction (DDI) and contraindication resources, revises unsafe predictions, and produces a traceable report. We evaluate SafeRx-Agent on MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2024) using both medication prediction accuracy and safety metrics. SafeRx-Agent outperforms traditional deep learning, LLM, and agentic baselines in prediction accuracy while reducing DDI and contraindication rates through explicit safety verification. Our key contributions are:

- We introduce the first fine-grained medication recommendation setting for predicting ATC-L4 code sets from EHRs, moving beyond the coarse ATC-L3 setting used in prior benchmarks, and show that this coarse aggregation systematically overestimates measured safety risk.
- We propose SafeRx-Agent, a medication recommendation multi-agent framework that combines specialty-aware generation, evidence grounding, safety-aware revision, and traceable reporting in a unified workflow.
- We introduce a knowledge-grounded safety verifier that detects the risks of drug interactions and contraindications, revises unsafe candidates, and produces traceable medication reports.
- We evaluate SafeRx-Agent on two real-world EHR datasets, showing improved fine-grained prediction accuracy, lower safety risks, and predicted set sizes closer to the ground truth.

2 Related Work

Supervised Medication Recommendation. Supervised methods cast medication recommendation as multi-label prediction over structured EHR codes, with GAMENet (Shang et al., 2019) introducing graph-augmented memory and DDI-aware

decoding, and follow-ups adding molecular structure (Yang et al., 2021, 2023), copy-generate decoding (Wu et al., 2022), and rare-drug or cold-start training (Zhao et al., 2024; Kuang and Xie, 2024). These models operate on coarse ATC-L3 vocabularies and encode safety implicitly in the loss, but fail to perform fine-grained medication prediction.

LLMs for Medication Recommendation. Direct prompting of general or medical LLMs (Zhang et al., 2024; Christophe et al., 2024; Pal and Sankarasubbu, 2024; Chen et al., 2025; Garcia-Gasulla et al., 2026) leverages richer textual context than models based on structured EHR codes, but lacks built-in safety verification (Hager et al., 2024; Asgari et al., 2025; Farrag et al., 2026). Fine-tuning approaches such as LAMO (Zhao et al., 2025), FLAME (Fan et al., 2025), and LEADER (Liu et al., 2025) address safety via losses, rewards, or distillation, but require task-specific training and remain tied to fixed backbones.

Multi-Agent Frameworks for Clinical Decision Support. Multi-agent LLM systems distribute clinical reasoning across coordinated roles for medical QA (Tang et al., 2024; Kim et al., 2024) and rare-disease diagnosis and treatment (Chen et al., 2026). However, these frameworks target medical QA and treatment discussion rather than closed-vocabulary medication prediction with knowledge-grounded verification. SafeRx-Agent differs from prior work on both axes: at the task level it predicts ATC-L4 rather than ATC-L3 categories, and at the method level it moves beyond end-to-end prediction that encode safety implicitly in the training loss, instead using specialty-conditioned generation grounded in medication indication evidence, cross-specialty reconciliation, and explicit case-level adjudication of DDIs and drug–diagnosis contraindications, without task-specific training.

3 Problem Formulation and Knowledge Resources

3.1 Problem Formulation

Let a patient record be a temporally ordered sequence of ICU visits $(v_1, \dots, v_T)$, where each visit $v_t = (\mathcal{D}_t, \mathcal{P}_t, \mathcal{M}_t)$ contains diagnoses, procedures, and medications. Diagnoses are represented by ICD-CM codes, procedures by ICD-PCS codes, and medications by ATC-L4 codes. Given the past visits and the current diagnoses and procedures, the task is to predict the medications prescribed at the

current visit. The input for visit T is

$$X_T = (\{(\mathcal{D}_t, \mathcal{P}_t, \mathcal{M}_t)\}_{t < T}, \mathcal{D}_T, \mathcal{P}_T).$$

The ground-truth medication set is $\mathcal{M}_T \subseteq \mathcal{V}_{\text{med}}$, and the model outputs a predicted set $\hat{\mathcal{M}}_T \subseteq \mathcal{V}_{\text{med}}$, where $\mathcal{V}_{\text{med}}$ is the ATC-L4 medication vocabulary. We evaluate prediction quality by comparing $\hat{\mathcal{M}}_T$ with $\mathcal{M}_T$, and evaluate safety by measuring DDIs and contraindications in $\hat{\mathcal{M}}_T$.

3.2 Clinical Knowledge Resources

SafeRx-Agent uses external clinical knowledge for evidence-grounded medication generation and safety-aware revision. We organize these resources around three functions:

- **Standardization:** diagnosis and medication taxonomies align EHR concepts with the ATC-L4 medication vocabulary.
- **Grounding:** indication evidence links patient conditions to clinically relevant medications.
- **Safety checking:** drug interaction and contraindication resources identify co-prescription and disease-conditioned risks.

Table 6 (Appendix A) summarizes the resources used, together with details about preprocessing, identifier mapping, and matrix construction.

4 SafeRx-Agent

4.1 Overview

As shown in Figure 1, SafeRx-Agent predicts the medication set $\hat{\mathcal{M}}_T \subseteq \mathcal{V}_{\text{med}}$ for the target visit from the longitudinal patient record X_T . SafeRx-Agent uses the clinical knowledge resources in Table 6 to ground medication generation and verify safety risks. SafeRx-Agent uses a multi-agent design for two reasons. At the workflow level, fine-grained prescribing couples heterogeneous objectives, namely evidence retrieval, clinical coverage, global consistency, and safety, which interfere when a single pass must balance them simultaneously; staging them lets the system first establish broad coverage and then progressively enforce consistency and safety. At the prediction level, ICU patients present with multiple comorbidities whose treatments fall under different therapeutic domains, so specialty agents contribute complementary views of the same patient while the global critique prevents the system from simply taking their union. Section 6.3 quantifies both effects.

Algorithm 1: SafeRx-Agent Inference

Input: Patient record X_T ; expert panel $\mathcal{E}$; indication resource R ; safety matrices $\mathbf{M}_{\text{DDI}}$, $\mathbf{M}_{\text{Contra}}$.

Output: Predicted medication set $\hat{\mathcal{M}}_T$.

▷ **Multi-agent generation**

```

1  $A \leftarrow \text{ROUTE}(X_T, \mathcal{E}), A \subseteq \mathcal{E}$ 
2 foreach  $E_i \in A$  do
3    $(s_i, \rho_i) \leftarrow \text{SUMMARIZE}(E_i, X_T)$ 
4    $\mathcal{M}_i \leftarrow \text{GENERATE}(E_i, s_i, \rho_i, X_T, R)$ 
5  $\tilde{\mathcal{M}} \leftarrow \text{CRITIQUE}(\{\mathcal{M}_i, s_i, \rho_i\}_{i \in A}, X_T, R)$ 
  ▷ Safety verification
6  $\mathcal{F} \leftarrow \text{FINDFLAGS}(\tilde{\mathcal{M}}, X_T, \mathbf{M}_{\text{DDI}}, \mathbf{M}_{\text{Contra}})$ 
7  $\hat{\mathcal{M}}_T \leftarrow \tilde{\mathcal{M}}$ 
8 foreach  $f \in \mathcal{F}$  do
9   if  $\text{VERIFY}(f, \hat{\mathcal{M}}_T, X_T, R) = \text{REM}$  then
10     $\hat{\mathcal{M}}_T \leftarrow \hat{\mathcal{M}}_T \setminus \{m_f\}$ 
11 return  $\hat{\mathcal{M}}_T$ 

```

Algorithm 1 summarizes the workflow, and the following subsections define each operator. The prompt templates used by the LLM-based operators are provided in Appendix M, Figures 7–10.

4.2 Expert Panel

We construct a reusable expert panel $\mathcal{E} = \{E_1, \dots, E_K\} \cup \{E_{\text{univ}}\}$ before inference. The panel contains K specialty experts and one always-on universal supportive-care expert E_{univ} for common inpatient medications. We derive the K specialty experts empirically by clustering patient embeddings and mapping each cluster to an ICD chapter-level scope. Appendix B provides implementation details.

4.3 Multi-Agent Generation

ROUTE: sparse specialty activation. Using all experts for each case increases computation and may introduce irrelevant specialty information, so ROUTE activates only relevant experts by matching patient’s current and historical ICD codes against each expert’s scope, with E_{univ} always included:

- $A \leftarrow \text{ROUTE}(X_T, \mathcal{E}), A \subseteq \mathcal{E}$.

Routing is sparse, so the number of expert calls scales with case complexity rather than the full panel size $|\mathcal{E}|$.

SUMMARIZE: expert-specific patient summarization. The raw record X_T remains long and heterogeneous, with each specialty’s relevant evidence scattered among entries outside its scope. SUMMARIZE therefore extracts and organizes the specialty-relevant information in X_T into compact

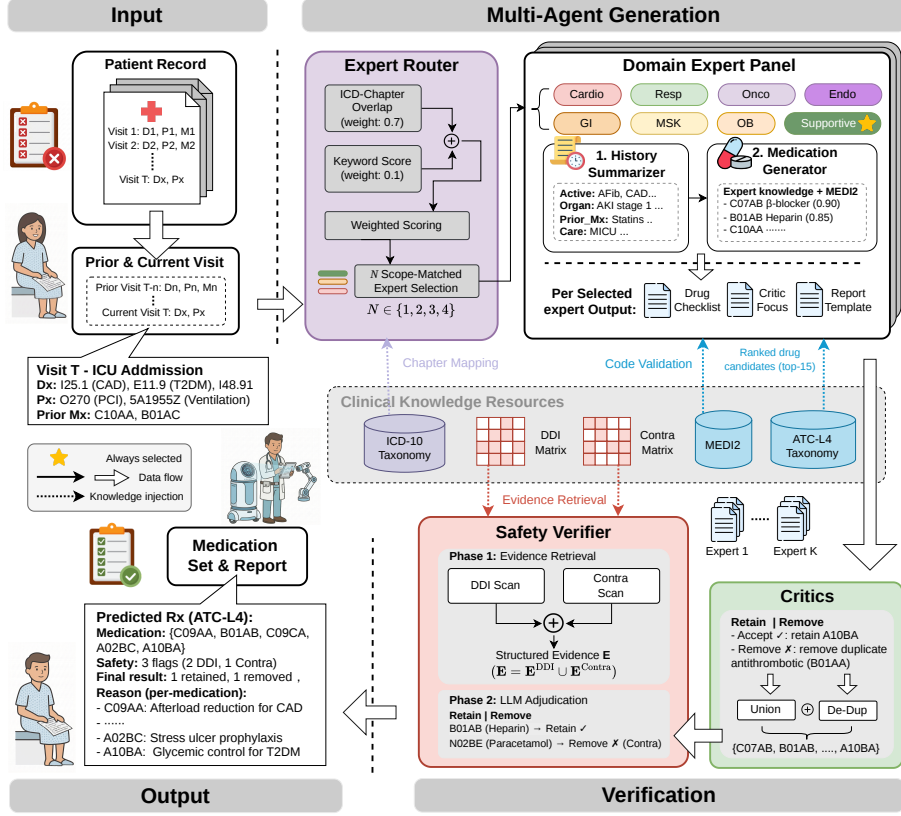


Figure 1: **Overview of SafeRx-Agent.** A patient record is routed via weighted ICD-chapter and keyword scoring to a sparse subset of specialty experts plus an always-on supportive-care expert. Each activated expert summarizes the patient record from its specialty scope and generates ATC-L4 medication candidates grounded in MEDI and the ATC taxonomy. A global Critique then judges and reconciles the expert proposals under the full patient context to produce a candidate set. A two-phase Safety Verifier retrieves DDI and contraindication evidence (M_{DDI} , M_{Contra}) and adjudicates each flag in context, yielding a traceable ATC-L4 prescription with per-medication rationales.

artifacts for each expert. For each activated expert E_i , SUMMARIZE converts the multi-visit record X_T into two expert-specific artifacts (Fig. 7):

- **Typed summary** s_i : a structured summary of medication-related evidence, including active problems, acute organ dysfunction, procedures, care context, medication-relevant risks, and prior prescriptions.
- **Visit rationales** ρ_i : structured explanations for prior visits that link diagnoses, procedures, and prescribed drugs, helping distinguish continued medications from transient treatments.

GENERATE: per-expert medication proposal.

For each activated expert E_i , GENERATE proposes medications based on the typed summary s_i , visit rationales ρ_i , and diagnosis-linked candidates retrieved from the MEDI resource R (Row 3, Table 6). It also consults the raw record X_T for details missing from s_i or ρ_i :

- $\mathcal{M}_i \leftarrow \text{GENERATE}(E_i, s_i, \rho_i, X_T, R)$.

CRITIQUE: global critique and reconciliation.

Since each expert provides specialty-specific medication proposals, the combined proposals may overlap, omit cross-specialty context, or include medications that are plausible within one specialty but weakly supported by the full patient context. CRITIQUE reviews the combined proposals using the expert-specific summaries s_i , visit rationales ρ_i , indication evidence R , and the full record X_T , then merges duplicate candidates and removes medications lacking sufficient patient-level support:

- $\tilde{\mathcal{M}} \leftarrow \text{CRITIQUE}(\{\mathcal{M}_i, s_i, \rho_i\}_{i \in A}, X_T, R)$.

4.4 Safety Verification

SafeRx-Agent verifies each generated candidate set because clinically plausible medications may still pose co-prescription or disease-conditioned safety risks. Safety verification takes the candidate set $\tilde{\mathcal{M}}$ as input for safety screening. It uses the DDI and contraindication matrices, M_{DDI} and M_{Contra} , derived from the safety resources in Table 6, rows 4

and 5. The verifier uses safety evidence and patient context to decide whether to retain or remove each flagged medication.

FINDFLAGS: matrix-based evidence collection.

FINDFLAGS scans $\tilde{\mathcal{M}}$ against $\mathbf{M}_{\text{DDI}}$ and $\mathbf{M}_{\text{Contra}}$ to collect a set of flagged items $\mathcal{F}$. Each flag is represented as $f = (m_f, \text{rel}_f, \text{evid}_f)$, where m_f is the flagged medication, $\text{rel}_f \in \{\text{DDI}, \text{CONTRA}\}$ is the triggering safety relation, and evid_f contains case-level evidence such as DDI degree and prior-prescription status.

VERIFY: case-specific adjudication. For each flagged item, VERIFY outputs an action $a_f \in \{\text{RET}, \text{REM}\}$ based on the flag f , the patient context X_T , and the indication resource R (Fig. 10). RET retains a flagged medication when the patient context supports its use, such as a previously tolerated DDI pair. REM removes a medication when the risk lacks patient-specific justification, such as a contraindication against an active diagnosis without prior exposure. The LLM is therefore not a source of safety knowledge: FINDFLAGS retrieves every relation deterministically, and VERIFY’s closed action space lets it neither introduce a medication outside $\tilde{\mathcal{M}}$ nor assert an undocumented relation, making it a constrained adjudicator rather than an open-ended prescriber. This preserves evidence-supported medication generation while removing candidates with unsupported safety risks. Appendix G audits its decisions against the intended patient-conditioned policy.

5 Experiment Design

5.1 Datasets and Preprocessing

We evaluate SafeRx-Agent on MIMIC-III and MIMIC-IV. Diagnoses and procedures are represented by ICD codes with textual descriptions; medications are normalized to ATC-L4 codes with within-visit duplicates removed. The prediction vocabulary $\mathcal{V}_{\text{med}}$ is the closed set of retained ATC-L4 codes; out-of-vocabulary predictions are discarded at evaluation time. All splits are patient-level. The MIMIC-IV cohort contains 1,686 patients, of which we hold out 100 for prompt development. Prompts are frozen afterwards, so neither the remaining 1,586-case MIMIC-IV evaluation set nor the 901-case MIMIC-III cohort, which is used as an external test set with prompts transferred without re-tuning, informs prompt design. Supervised baselines require training, so we additionally

partition the MIMIC-IV cohort into training, validation, and test splits, containing 1,124, 281, and 281 patients, and report their results on the held-out test split, whereas the training-free LLM-based methods are evaluated on their respective full evaluation set. No method is evaluated on data used for its own training or prompt development, and all methods share the same serialized input, vocabulary, and evaluation script (Table 10, Appendix C).

5.2 Baselines

We compare SafeRx-Agent against four baseline groups, isolating the effects of task-specific training, domain adaptation, general LLM capability, and agent workflow design. (1) **Traditional deep learning:** GAMENet (Shang et al., 2019), SafeDrug (Yang et al., 2021), DrugDoctor (Kuang and Xie, 2024), and KEHGCN (Zhang et al., 2026). (2) **Medical LLMs:** UltraMedical-70B (Zhang et al., 2024), Med42-v2-70B (Christophe et al., 2024), OpenBioLLM-70B (Pal and Sankarasubbu, 2024), Llama3.1-Aloe-Beta-70B (Garcia-Gasulla et al., 2026), and HuatuoGPT-o1 (Chen et al., 2025). (3) **General LLMs:** GPT-5.2, Claude Sonnet-4.6, Qwen3-32B (Yang et al., 2025), and Gemma3-27B-IT (Team et al., 2025). (4) **Agent-based baselines:** RareAgents (Chen et al., 2026) adapted for ATC-L4 code generation, and General Agent with a single ICU medication expert.

5.3 Implementation Details

Supervised baselines are trained on the patient-level splits with an output head over the ATC-L4 vocabulary and hyperparameters tuned on the validation set. All LLM-based methods share the same serialized EHR input and closed ATC-L4 vocabulary. The general-agent baseline replaces SafeRx-Agent’s routed specialty experts with one generalist agent, and RareAgents is adapted to ATC-L4 generation. Full baseline implementations, prompts, and inference settings are provided in Appendix D.

5.4 Evaluation Metrics

We evaluate each method along two axes: medication-set prediction quality and safety. For prediction quality, we report Jaccard, precision, recall, F1, and the average number of predicted ATC-L4 codes. For safety, we report GT-normalized DDI rates (DDI-B and DDI-W) and contraindication rates (Contra-B and Contra-W), where DDI-W and Contra-W weight each safety relation by its

observed frequency. Lower DDI and contraindication rates indicate safer predictions. Full metric definitions are provided in Appendix E.

6 Results

6.1 Quantitative Results

Table 1 reports medication recommendation performance on MIMIC-IV and MIMIC-III. SafeRx-Agent consistently outperforms direct prompting and generic agent baselines across all backbones, achieving the best pre-filter F1 on MIMIC-IV for every model tested. The safety filter trades a small drop in recall for higher precision and smaller predicted sets, as expected from a verifier that removes unsupported candidates. Without task-specific training, SafeRx-Agent also matches or exceeds supervised baselines on F1 and Jaccard, although the supervised rows are computed on their own held-out test split (Section 5.1) and are therefore best read as reference operating points. MIMIC-III shows the same pattern, confirming that the workflow generalizes beyond MIMIC-IV. Case-level paired bootstrap (5,000 resamples) confirms the margin over the strongest agent baseline, +0.039 F1 on MIMIC-IV ($p = 0.019$) and +0.116 on MIMIC-III ($p < 0.005$), an order of magnitude above the three-seed F1 standard deviations (Appendix F).

6.2 Drug Safety Analysis

We evaluate safety by comparing each predicted ATC-L4 set with the DDI and contraindication matrices constructed from external resources (Table 2). Deep learning baselines and unfiltered LLM agents generate larger medication sets with higher DDI and contraindication rates, so they cannot reliably control safety risks without explicit verification. Our safety filter consistently reduces both rates while keeping the predicted set size closer to the ground truth, with the largest reductions for GPT-5.2 and Gemma3-27B-IT. The reduction is statistically reliable: the paired-bootstrap DDI-W change from removing the verifier is -0.290 on MIMIC-IV (95% CI $[-0.491, -0.137]$) and -0.401 on MIMIC-III (95% CI $[-0.569, -0.140]$), both $p < 0.001$ (Appendix F). We also note that the safety matrices are used only by FINDFLAGS to surface candidate risks and never determine the action: VERIFY adjudicates each flag against the patient record and may retain a flagged medication, so these metrics measure selective enforcement of

documented safety knowledge rather than reproduction of a matrix lookup. Overall, these findings suggest that reliable fine-grained medication recommendation requires explicit safety verification instead of relying on a supervised learning loss or LLM generation alone.

6.3 Ablation Studies

To quantify the marginal contribution of each component, we ablate SafeRx-Agent by removing one component at a time while keeping all other settings fixed, on MIMIC-IV with Gemma3-27B-IT (Table 3); all subsequent analyses use this setting. Replacing the specialty panel with a general agent causes the largest F1 drop (-0.057) and enlarges the predicted set (10.7 vs. 9.65), confirming that specialty routing constrains each expert to its scope. Removing the critique hurts precision most ($0.379 \rightarrow 0.322$) while recall stays flat, consistent with its role as a pruning stage. Removing the medical resource instead hurts recall most ($0.398 \rightarrow 0.325$), indicating that retrieved evidence helps recover medications not surfaced by the LLM alone.

6.4 Controlled Analysis of Architecture, Budget, and Set Size

Two confounders could explain SafeRx-Agent’s gains without any contribution from its architecture: privileged access to external knowledge, and a larger inference-compute budget. A third could explain its safety gains: predicting fewer medications. We isolate all three on MIMIC-IV with Gemma3-27B-IT, holding the patient input, closed vocabulary, backbone, and decoding settings fixed.

Knowledge and inference budget. Table 4 adds a MEDI-matched direct-prompting variant and a call-matched general agent, which repeats the general agent’s generation and critique stack over independent self-consistency passes for 10.0 LLM calls per case, above SafeRx-Agent’s 8.4 generation calls and 9.4 total calls. We match calls rather than tokens because SafeRx-Agent processes short specialty-scoped contexts while each generalist pass processes the full record, so token equality would require truncation or padding that changes the methods. Giving direct prompting the same indication resource changes F1 by only $+0.002$ ($0.3174 \rightarrow 0.319$), and more than tripling the general agent’s compute by $+0.004$ ($0.3313 \rightarrow 0.335$) while lowering Jaccard; SafeRx-Agent exceeds the

Table 1: Prediction performance across MIMIC-IV and MIMIC-III. Within each LLM backbone group, **bold** marks the best and underline marks the second-best (excluding the Avg. #Pred column). Avg. #Pred reports the average number of predicted medications per case and is compared with the average ground-truth size. Supervised baselines are evaluated on their held-out test split and LLM-based methods on the full evaluation set (Section 5.1), so the two blocks are reference operating points rather than paired comparisons.

Method	MIMIC-IV					MIMIC-III				
	Precision	Recall	F1	Jaccard	Avg. #Pred	Precision	Recall	F1	Jaccard	Avg. #Pred
<i>Deep learning medication recommendation baselines</i>										
GAMENet	0.2510	0.5062	0.3356	0.2171	23.92	0.1521	0.3684	0.2153	0.1834	18.16
SafeDrug	0.2639	0.5280	0.3519	0.2286	16.53	0.1638	0.4227	0.2361	0.1914	13.29
DrugDoctor	0.2796	0.5190	0.3634	0.2369	17.17	0.1551	0.3915	0.2232	0.1909	13.88
KEHGCN	0.2980	0.5493	0.3864	0.2548	21.31	0.1634	0.4284	0.2365	0.2018	16.75
<i>Open-source Medical LLM baselines (direct prompting)</i>										
UltraMedical-70B	0.2992	0.2402	0.2665	0.1493	7.36	0.1254	0.0999	0.1112	0.0486	4.27
Med42-v2-70B	0.2414	0.2220	0.2313	0.1273	8.44	0.1293	0.1245	0.1269	0.0689	5.16
Llama3.1-Aloe-Beta-70B	0.3234	0.2370	0.2735	0.1572	6.72	0.1499	0.1098	0.1268	0.0708	3.92
HuatuoGPT-o1	0.3100	0.1417	0.1945	0.1085	4.19	0.1476	0.1162	0.1300	0.0715	4.22
OpenBioLLM-70B	0.4104	0.3691	0.3887	0.2481	8.25	0.1699	0.0754	0.1045	0.0529	2.38
<i>Closed-source LLM: GPT-5.2</i>										
Direct Prompting	0.3528	0.4447	0.3935	0.2535	11.62	0.1891	0.3383	0.2426	0.1424	9.57
General Agent	0.4271	0.3562	0.3884	0.2475	8.50	0.2087	<u>0.3215</u>	0.2531	0.1526	8.24
RareAgents	0.3902	0.4072	0.3985	0.2573	9.61	0.1586	0.2019	0.1776	0.1031	6.81
Ours (w/o safety filter)	0.4489	0.3837	0.4138	0.2570	8.71	0.2991	0.2654	0.2812	0.1702	4.70
Ours (w/ safety filter)	0.5210	0.3160	0.3934	0.2631	6.18	0.3145	0.2277	<u>0.2641</u>	<u>0.1620</u>	3.85
<i>Closed-source LLM: Claude Sonnet-4.6</i>										
Direct Prompting	0.2650	0.5662	0.3610	0.2100	19.70	0.1328	0.4305	0.2029	0.1103	17.25
General Agent	0.3016	0.5152	0.3804	0.2229	17.41	0.1513	<u>0.4323</u>	0.2240	0.1250	15.20
RareAgents	0.3656	0.4343	0.3970	0.2531	10.94	0.1260	0.2152	0.1590	0.0873	9.56
Ours (w/o safety filter)	0.3557	<u>0.5515</u>	0.4325	0.2650	15.80	<u>0.1950</u>	0.4711	0.2758	0.1586	12.37
Ours (w/ safety filter)	<u>0.3611</u>	0.5103	<u>0.4229</u>	<u>0.2617</u>	14.40	0.1971	0.4125	<u>0.2667</u>	<u>0.1528</u>	10.72
<i>Open-source LLM: Qwen3-32B</i>										
Direct Prompting	0.3452	0.3145	0.3292	0.2017	8.36	0.1381	0.1086	0.1216	0.0616	4.22
General Agent	0.3717	0.2978	0.3307	0.2030	6.48	0.1828	0.1067	0.1350	0.0700	3.14
RareAgents	<u>0.3724</u>	0.2931	0.3280	0.2011	7.24	0.1612	0.0806	0.1074	0.0552	2.74
Ours (w/o safety filter)	0.3939	<u>0.3277</u>	0.3578	0.2276	7.38	0.2899	0.2712	0.2802	0.1569	4.53
Ours (w/ safety filter)	0.3576	0.3418	<u>0.3495</u>	<u>0.2212</u>	6.51	<u>0.2384</u>	<u>0.2533</u>	<u>0.2456</u>	<u>0.1265</u>	3.44
<i>Open-source LLM: Gemma3-27B-IT</i>										
Direct Prompting	0.2795	<u>0.3673</u>	0.3174	0.1856	12.06	0.0986	0.1477	0.1183	0.0634	8.03
General Agent	0.3073	0.3595	0.3313	0.1984	10.7	0.1349	0.1411	0.1379	0.0723	5.63
RareAgents	0.3677	0.3050	0.3334	0.2006	7.61	0.1041	0.0773	0.0887	0.0462	3.98
Ours (w/o safety filter)	0.3789	0.3978	0.3882	0.2446	9.65	<u>0.2289</u>	0.3121	0.2641	0.1537	7.30
Ours (w/ safety filter)	0.4374	0.3239	<u>0.3722</u>	0.2512	6.79	0.2479	<u>0.2567</u>	<u>0.2522</u>	<u>0.1146</u>	5.39

call-matched general agent by +0.053 F1 (95% CI [0.018, 0.103], $p = 0.006$). The identical verifier lowers documented risk for both stacks, yet SafeRx-Agent stays safer afterwards (DDI-B 0.3695 vs. 0.5970; Appendix G), so the advantage is not the verifier’s alone. The gains therefore track the specialty decomposition rather than knowledge access or inference budget.

Prediction-set size. Because verification shrinks the predicted set, we ask whether its safety gain is merely a size effect. For each case we delete $k = |\hat{\mathcal{M}}^{\text{pre}}| - |\hat{\mathcal{M}}^{\text{VERIFY}}|$ medications from the pre-verification set, matching VERIFY’s per-case size exactly, and vary only the deletion rule (Table 16, Appendix G). The four control rules are patient-agnostic; only VERIFY conditions each decision on the individual patient. At identical per-case set size, VERIFY attains the lowest DDI and contraindication rates by a wide margin, 0.3695 versus 0.5479 DDI-B for the strongest control, so the reduction is

risk-targeted not a by-product of predicting less.

6.5 Diagnostic Analysis

We analyze how the critique stage reshapes the proposed medication set by comparing pre- and post-critique predictions. Figure 5a shows that critique reduces false positives from 9.51 to 6.01 per case at a cost of only 0.46 true positives, acting as a precision-oriented pruning stage, and Figure 5b reveals why: of the 6,625 removed medications, 6,373 (96.2%) were proposed by a single expert, while medications corroborated by multiple specialists were rarely pruned. The critique therefore targets candidates lacking independent corroboration rather than indiscriminately shrinking the set. Appendices H and N report supplementary diagnostics and qualitative case studies.

6.6 Expert Analysis

To test whether individual experts provide marginal value beyond the panel as a whole, we conduct a

Table 2: Safety analysis of predicted medication sets across MIMIC-IV and MIMIC-III. All safety metrics are lower-is-better. Within each LLM backbone group, **bold** marks the best and underline marks the second-best in each column (excluding the Ground Truth row and the Avg. #Pred column). DDI-B and Contra-B report the fraction of cases with at least one DDI or contraindication, while DDI-W and Contra-W report their frequency-weighted rates. Avg. #Pred is the average number of predicted ATC codes per case. As in Table 1, supervised rows use the supervised test split and LLM-based rows the full evaluation set, to which Ground Truth corresponds.

Method	MIMIC-IV ($\downarrow$ except #Pred)					MIMIC-III ($\downarrow$ except #Pred)				
	DDI-B	DDI-W	Contra-B	Contra-W	Avg. #Pred	DDI-B	DDI-W	Contra-B	Contra-W	Avg. #Pred
Ground Truth	0.3013	0.1843	0.0043	0.0011	9.17	0.2460	0.1507	0.0023	0.0007	5.36
<i>Deep learning medication recommendation baselines</i>										
GAMENet	0.7516	0.6023	0.0165	0.0074	23.92	0.6238	0.5848	0.0081	0.0030	18.16
SafeDrug	0.5869	0.4326	0.0083	0.0035	16.53	0.4055	0.3268	0.0053	0.0015	13.29
DrugDoctor	0.7010	0.5867	0.0157	0.0049	17.17	0.5727	0.5058	0.0075	0.0022	13.88
KEHGCN	0.6134	0.4653	0.0072	0.0031	21.31	0.4351	0.3044	0.0042	0.0013	16.75
<i>Medical LLM baselines (direct prompting)</i>										
UltraMedical-70B	0.4033	0.3132	0.0058	0.0016	7.36	0.3356	0.2644	0.0043	0.0015	4.27
Med42-v2-70B	0.5477	0.4199	0.0075	0.0024	8.44	0.5206	0.3922	0.0063	0.0021	5.16
Llama3.1-Aloe-Beta-70B	0.4245	0.3092	0.0056	0.0015	6.72	0.3840	0.2685	0.0056	0.0016	3.92
HuatuoGPT-o1	0.2330	0.1592	0.0035	0.0011	4.19	0.3503	0.2376	0.0046	0.0015	4.22
OpenBioLLM-70B	0.3878	0.2948	0.0054	0.0013	8.25	0.2453	0.1732	0.0028	0.0009	2.38
<i>Closed-source LLM: GPT-5.2</i>										
Direct Prompting	0.7218	0.5914	0.0067	0.0014	11.62	0.9022	0.8360	0.0010	0.0001	9.57
General Agent	0.4896	0.3902	0.0060	0.0013	8.50	0.8686	0.7773	0.0010	0.0004	8.24
RareAgents	0.4980	0.3650	0.0030	0.0006	9.61	0.6101	0.5214	0.0004	0.0001	6.81
Ours (w/o safety filter)	0.5340	0.4126	0.0078	0.0017	8.71	0.5765	0.4833	0.0050	0.0007	4.70
Ours (w/ safety filter)	0.2819	0.2115	0.0028	0.0007	6.18	0.4386	0.3572	0.0000	0.0000	3.85
<i>Closed-source LLM: Claude Sonnet-4.6</i>										
Direct Prompting	0.9408	0.8668	0.0106	0.0026	19.70	0.9875	0.9643	0.0140	0.0031	17.25
General Agent	0.8951	0.8061	0.0106	0.0032	17.41	0.9886	0.9582	0.0134	0.0029	15.20
RareAgents	0.6162	0.4952	0.0043	0.0010	10.94	0.8007	0.6990	0.0000	0.0000	9.56
Ours (w/o safety filter)	0.9096	0.8283	0.0076	0.0013	15.80	0.9679	0.9403	0.0048	0.0012	12.37
Ours (w/ safety filter)	0.7981	0.7068	0.0030	0.0009	14.40	0.9376	0.8877	0.0006	0.0003	10.72
<i>Open-source LLM: Qwen3-32B</i>										
Direct Prompting	0.5318	0.4095	0.0074	0.0018	8.36	0.4243	0.3168	0.0038	0.0011	4.22
General Agent	0.4100	0.3185	0.0050	0.0013	6.48	0.3249	0.2389	0.0030	0.0007	3.14
RareAgents	0.3959	0.3107	0.0077	0.0022	7.24	0.1730	0.1154	0.0046	0.0013	2.74
Ours (w/o safety filter)	0.5262	0.4177	0.0044	0.0011	7.38	0.7396	0.4698	0.0090	0.0023	4.53
Ours (w/ safety filter)	0.3585	0.3758	0.0037	0.0007	6.51	0.3927	0.3513	0.0000	0.0000	3.44
<i>Open-source LLM: Gemma3-27B-IT</i>										
Direct Prompting	0.6960	0.5747	0.0095	0.0022	12.06	0.7643	0.6441	0.0061	0.0015	8.03
General Agent	0.6620	0.5530	0.0088	0.0022	10.7	0.6296	0.5304	0.0065	0.0019	5.63
RareAgents	0.4753	0.3624	0.0069	0.0017	7.61	0.4876	0.3828	0.0028	0.0010	3.98
Ours (w/o safety filter)	0.6683	0.5761	0.0086	0.0017	9.65	0.8170	0.7268	0.0062	0.0013	7.30
Ours (w/ safety filter)	0.3695	0.2865	0.0030	0.0007	6.79	0.3650	0.3258	0.0000	0.0000	5.39

leave-one-expert-out (LOO) analysis on activation-defined subgroups. For each expert, we compare subgroup F1 before and after removing only that expert. Figure 5c shows that every activated expert contributes positively, with F1 drops ranging from +0.101 (RESP) to +0.232 (SUP); the supportive-care expert contributes the most, consistent with its broad activation, and even the least impactful expert yields a meaningful gain, so no activated expert is redundant. Routing is multi-label, activating 2–3 specialty experts per case alongside the always-on supportive-care expert, so no unique gold specialty set exists and a conventional routing error rate is undefined; the relevant question is whether non-routed experts would recover medications that sparse routing misses. Activating the full panel instead raises recall by only 0.030 while lowering precision and F1, raising DDI-B

(0.6683→0.8698) and Contra-B (0.0086→0.0098), and doubling generation-stage LLM calls (17.0 vs. 8.4; Appendix J), so indiscriminate activation mainly injects off-scope candidates that routing already excludes. Appendix I reports activation sparsity and per-expert TP contributions.

6.7 Medication Granularity and Safety Evaluation

To quantify how medication granularity affects safety evaluation, we hold the medication sets fixed and vary only the representation. We compute binary DDI and contraindication rates directly from the ground-truth ATC-L4 medication sets, then collapse each ATC-L4 code to its ATC-L3 parent and recompute the same metrics at the ATC-L3 level. Because a documented relation between two chemical subgroups is inherited by their pharmacological

Table 3: **Prediction ablations.** Each variant removes one component while keeping other settings fixed. **Red** marks decreases. #Pred deltas are shown in black since smaller or larger sets are not directly better or worse.

Method	Precision	Recall	F1	Jaccard	#Pred
Full SafeRx-Agent	0.3789	0.3978	0.3882	0.2446	9.65
w/o Specialty Experts	0.3073 (−0.0716)	0.3595 (−0.0383)	0.3313 (−0.0569)	0.1984 (−0.0462)	10.7 (+1.05)
w/o Summarizer	0.3564 (−0.0225)	0.3795 (−0.0183)	0.3679 (−0.0203)	0.2272 (−0.0174)	9.28 (−0.37)
w/o Medical Resource	0.3699 (−0.0090)	0.3249 (−0.0729)	0.3673 (−0.0209)	0.2231 (−0.0215)	9.57 (−0.08)
w/o Critique	0.3225 (−0.0564)	0.3849 (−0.0129)	0.3589 (−0.0293)	0.2370 (−0.0076)	9.71 (+0.06)

Table 4: **Controlled comparison under matched knowledge access and inference budget** (MIMIC-IV, Gemma3-27B-IT, before safety verification). They share patient input, vocabulary, backbone, and decoding settings. Rows 0, 2 and 4 restate Table 1; rows 1 and 3 are controls. Calls = generation-stack LLM calls per case, Rt. = routing, Cr. = critique, GA = general agent.

#	Variant	MEDI	Rt.	Cr.	Calls	Prec.	Rec.	F1
0	Direct	—	—	—	1.0	0.280	0.367	0.317
1	Direct + MEDI	✓	—	—	2.0	0.283	0.365	0.319
2	GA + MEDI	✓	—	✓	3.0	0.307	0.360	0.331
3	Call-matched GA	✓	—	✓	10.0	0.305	0.371	0.335
4	SafeRx-Agent	✓	✓	✓	8.4	0.379	0.398	0.388

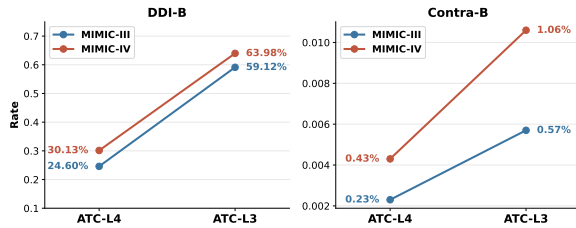


Figure 2: Binary DDI and contraindication rates for the same ground-truth medication sets evaluated under the ATC-L4 and ATC-L3 vocabularies. Collapsing L4 chemical subgroups to L3 pharmacological parents causes parent pairs to inherit documented relations from their children, inflating measured risk on both datasets.

parents, every sibling subgroup pair beneath those parents is flagged as well, even when no relation is recorded for it. Figure 2 shows the resulting inflation. On MIMIC-III, DDI-B rises from 24.60% to 59.12% and Contra-B from 0.23% to 0.57%; on MIMIC-IV, DDI-B rises from 30.13% to 63.98% and Contra-B from 0.43% to 1.06%. Coarse ATC-L3 evaluation therefore assigns clinically real prescriptions more than twice their documented interaction burden, which supports fine-grained recommendation at ATC-L4.

6.8 Efficiency Analysis

We evaluate the inference efficiency of SafeRx-Agent on MIMIC-IV in terms of average LLM calls, input and output tokens, and per-case latency (Table 5). Despite its evidence-grounded multi-agent workflow, SafeRx-Agent uses fewer LLM calls than RareAgents (9.4 vs. 10.1), suggesting

Table 5: **Efficiency analysis.** We report average LLM calls, token usage, and wall-clock latency per case.

Method	#Calls	In Tok.	Out Tok.	Sec/case
Direct Prompting	1.0	885	517	9.6
General Agent	3.0	4,928	788	10.2
RareAgents	10.1	13,478	1,710	13.3
SafeRx-Agent w/o safety filter	8.4	24,260	5,988	27.7
SafeRx-Agent w/ safety filter	9.4	24,847	6,070	29.5

that sparse routing avoids unnecessary expert calls, and the safety filter is lightweight, adding one LLM call and raising latency from 27.7 to 29.5 seconds per case. SafeRx-Agent nevertheless occupies a different operating point from lightweight supervised models, which run one forward pass and are preferable when serving latency dominates, but need dataset-specific training and give fixed-function predictions without inference-time evidence, traceable decisions, or case-specific safety adjudication. SafeRx-Agent is training-free at deployment, runs on multiple backbones, transfers across datasets without prompt re-tuning, and spends its extra inference compute on grounding, traceability, and explicit safety control rather than on latency (Appendices L and K).

7 Conclusion

We presented SafeRx-Agent, a knowledge-grounded multi-agent framework for safe and explainable medication recommendation, together with a fine-grained ATC-L4 setting that preserves subgroup distinctions hidden by coarser benchmarks. SafeRx-Agent routes each case to multiple specialty experts, grounds generation in patient context and indication evidence, and verifies safety against DDI and contraindication resources. On MIMIC-III and MIMIC-IV it improves prediction accuracy over deep learning, LLM, and agentic baselines while reducing safety risks and controlling set size, and controlled comparisons show these gains are not explained by knowledge access, inference budget, or smaller prediction sets. We hope this work encourage more verifiable agentic systems for high-stakes clinical NLP.

Limitations

SafeRx-Agent measures medication safety against external resources, so its safety claims concern *documented* risk: relations absent from TWOSIDES or openFDA are treated as safe by default, making the reported rates lower bounds (Appendices A.5 and E). The verifier draws on the same resources as the evaluation, following established practice in this literature, where interaction knowledge serves both as a training or inference-time signal and as an evaluation target. Because FINDFLAGS only surfaces candidate risks and VERIFY adjudicates each flag against the patient record (Section 6.2), our results show selective enforcement of documented safety knowledge; Appendix G audits this behavior, and validating it against blinded pharmacist review remains future work.

Our scope is ATC-L4 medication-set recommendation. ATC-L4 denotes chemical subgroups rather than ATC-L5 active ingredients, and we do not model dose, route, frequency, or duration, which depend on temporally richer context than our encounter-level input provides. Our claims are accordingly limited to ATC-L4 medication sets and to documented interactions and contraindications at that resolution. Our evaluation is also retrospective and offline; deployment in real clinical workflows would require prospective validation and physician oversight to ensure patient safety.

Acknowledgments

This work was supported in part by the NVIDIA Academic Grant Program through an award to Ziyang Song at Ohio University. We thank NVIDIA for providing resources that supported this research.

References

- Zafar Ali, Yi Huang, Irfan Ullah, Junlan Feng, Chao Deng, Nimbeshaho Thierry, Asad Khan, Asim Ullah Jan, Xiaoli Shen, Wu Rui, and Guilin Qi. 2023. Deep learning for medication recommendation: A systematic survey. *Data Intell.*, 5(2):303–354.
- Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. [A framework to assess clinical safety and hallucination rates of llms for medical text summarisation](#). *npj Digital Medicine*, 8:274.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2026. [Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. [Med42-v2: A suite of clinical llms](#). *Preprint*, arXiv:2408.06142.
- Chenxiao Fan, Chongming Gao, Wentao Shi, Yaxin Gong, Zhao Zihao, and Fuli Feng. 2025. [Fine-grained list-wise alignment for generative medication recommendation](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 48037–48061. Curran Associates, Inc.
- Ahmed N. Farrag, Amany El-Zeiny, and Amani M. Ali. 2026. [Evaluating large language models for pharmacotherapy simulations: a mixed-methods study](#). *npj Digital Medicine*, 9:355.
- Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Torrens, Daniel Hinos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, and 1 others. 2026. The aloe family recipe for open and specialized healthcare llms. *npj Digital Medicine*.
- Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. [Evaluation and mitigation of the limitations of large language models in clinical decision-making](#). *Nature Medicine*, 30:2613–2622.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. [MIMIC-IV](#). *PhysioNet*. Version 3.1.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2016. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.
- Taha A Kass-Hout, Zhiheng Xu, Matthew Mohebbi, Hans Nelsen, Adam Baker, Jonathan Levine, Elaine Johanson, and Rosalie A Bright. 2016. OpenFDA: an innovative platform providing access to a wealth of FDA’s publicly available data. *J. Am. Med. Inform. Assoc.*, 23(3):596–600.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. [Mdagents: An adaptive collaboration of llms for medical decision-making](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

- Yabin Kuang and Minzhu Xie. 2024. [Drugdoctor: enhancing drug recommendation in cold-start scenario via visit-level representation learning and training](#). *Briefings in Bioinformatics*, 25(6):bbae464.
- Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2025. [Large language model distilling medication recommendation model](#). *Preprint*, arXiv:2402.02803.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. [Gamenet: Graph augmented memory networks for recommending medication combination](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1126–1133.
- Ziyang Song, Yuanyi Hu, Aman Verma, David L. Buckeridge, and Yue Li. 2022. [Automatic phenotyping by a seed-guided topic model](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4713–4723, New York, NY, USA. Association for Computing Machinery.
- Ziyang Song, Qincheng Lu, Hao Xu, He Zhu, David Buckeridge, and Yue Li. 2024. [Timelygpt: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare](#). In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '24, New York, NY, USA. Association for Computing Machinery.
- Ziyang Song, Qincheng Lu, He Zhu, David Buckeridge, and Yue Li. 2026. [Trajgpt: Irregular time-series representation learning of health trajectory](#). *IEEE Journal of Biomedical and Health Informatics*, 30(5):3888–3899.
- Steven J Steindel. 2010. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J. Am. Med. Inform. Assoc.*, 17(3):274–282.
- 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, Bangkok, Thailand. Association for Computational Linguistics.
- Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. 2012. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.*, 4(125):125ra31.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Ruilin Wang, Bo-Hong Wang, Elizabeth Kourbatski, Jun Bai, Hegang Chen, Ziyang Song, Gilles Boire, Marie Hudson, and Yue Li. 2026. [Doctoragents: an agentic framework to iteratively refine automl pipeline for small clinical temporal data](#). *Preprint*, arXiv:2608.05375.
- Wei-Qi Wei, Robert M Cronin, Hua Xu, Thomas A Lasko, Lisa Bastarache, and Joshua C Denny. 2013. Development and evaluation of an ensemble resource linking medications to their indications. *J. Am. Med. Inform. Assoc.*, 20(5):954–961.
- WHO Collaborating Centre for Drug Statistics Methodology. 2026. [Guidelines for ATC classification and DDD assignment](#). Oslo.
- Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. [Conditional generation net for medication recommendation](#). In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 935–945, New York, NY, USA. Association for Computing Machinery.
- Jiabao Xu, Xuefeng Xi, Jie Chen, Victor S Sheng, Jieming Ma, and Zhiming Cui. 2022. A survey of deep learning for electronic health records. *Applied Sciences*, 12(22):11709.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. [Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3735–3741. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- 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*, WWW '23, page 4075–4085, New York, NY, USA. Association for Computing Machinery.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jin-Fang Hu,

Zhiyuan Liu, and Bowen Zhou. 2024. [Ultramedical: Building specialized generalists in biomedicine](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 26045–26081. Curran Associates, Inc.

Zihan Zhang, Hongzhi Liu, Xiaoshuang Guo, Tianqi Sun, and Zhonghai Wu. 2026. Knowledge-enhanced explainable HyperGraph convolution network for medication recommendation. *Proc. Conf. AAAI Artif. Intell.*, 40(19):16424–16432.

Zihao Zhao, Chenxiao Fan, Junlong Liu, Zheng Wang, Xiangnan He, Chongming Gao, Juan Li, and Fuli Feng. 2025. [Fine-grained alignment of large language models for general medication recommendation without overprescription](#). *Preprint*, arXiv:2503.03687.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. [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*, SIGIR '24, page 533–542, New York, NY, USA. Association for Computing Machinery.

A Details About Clinical Domain Knowledge

A.1 Construction of Medical Ontologies

We organize diagnosis and medication concepts into hierarchical ontologies to align EHR codes with external clinical knowledge. Diagnosis codes follow the ICD taxonomy based on CCS hierarchy, and medication codes follow the ATC taxonomy. Each code is assigned to a unique path from a broad category to a fine-grained leaf code. This path records the semantic lineage of the code and provides a consistent representation for evidence alignment, medication vocabulary mapping, and safety evaluation.

Diagnosis. Diagnostic concepts are organized using the CCS framework developed by the Agency for Healthcare Research and Quality (AHRQ), which groups ICD-9-CM and ICD-10-CM diagnosis codes into clinically coherent categories at multiple levels of granularity. We construct a five-level hierarchy from CCS major categories down to decimal-level ICD codes, ensuring that each ICD code maps to a unique hierarchical path.

Medication. Medication concepts are organized using the ATC classification system, which defines a five-level hierarchy from anatomical main groups down to chemical substances and encodes pharmacological relationships among drugs.

Per-level concept counts for both taxonomies are reported in Table 7.

A.2 Medication Indication

MEDI is an ensemble medication indication resource for primary and secondary uses of EHR data. MEDI was created based on multiple commonly used medication resources (RxNorm, MedlinePlus, SIDER 4.1, Mayo Clinic, WebMD, and Wikipedia) and by leveraging both ontology and NLP techniques. The current release of MEDI contains 3,031 medications and 186,064 indications, including both ICD-9-CM and ICD-10-CM indication pairs.

A.3 Drug-Drug Interactions

DDIs describe medication pairs with known co-prescription risks. We construct DDI resources from the TWOSIDES knowledge base (Tatonetti et al., 2012). TWOSIDES contains drug-pair interaction records with source drug identifiers. We map these identifiers to ATC-L4 medication codes and aggregate the mapped records over the ATC-L4 medication vocabulary used in this study. This mapping produces two DDI matrices. The binary DDI matrix records whether an ATC-L4 medication pair has at least one known interaction record in TWOSIDES. If any source drug pair mapped to the same ATC-L4 pair has a recorded interaction, the corresponding binary entry is set to one; otherwise, it is set to zero. This matrix is used to compute binary DDI rates. The weighted DDI matrix records the observed interaction frequency for each ATC-L4 medication pair. If multiple source drug pairs map to the same ATC-L4 pair, we sum their frequencies to obtain one pair-level weight. This matrix is used to compute frequency-weighted DDI rates, assigning larger penalties to medication pairs with more frequently observed interaction records. We symmetrize both matrices because DDI risks are defined over unordered medication pairs. We exclude diagonal entries when computing pairwise DDI rates.

A.4 Contraindications from openFDA

Contraindications specify clinical situations in which a medication may be unsafe because potential harms outweigh expected benefits. The openFDA Drug Labeling dataset provides programmatic access to FDA Structured Product Labeling (SPL) submissions for prescription and over-the-counter drugs (Kass-Hout et al., 2016). SPL

Table 6: **Clinical knowledge resources used in this work.** All medication-side resources are mapped to the ATC-L4 vocabulary. Contra. = Contraindication.

Resource	Derived representation	Used in SafeRx-Agent
ICD/CCS taxonomy (Steindel, 2010)	Diagnosis codes grouped into clinical chapters and subcategories	Expert panel & routing (Section 4.2 & 4.3)
ATC taxonomy (WHO Collaborating Centre for Drug Statistics Methodology, 2026)	Defines ATC-L4 prediction vocabulary $\mathcal{V}_{\text{med}}$	Throughout Section 4
MEDI (Wei et al., 2013)	Diagnosis–medication indication relation $R \subseteq \mathcal{V}_{\text{diag}} \times \mathcal{V}_{\text{med}}$	Generation & critique (Section 4.3)
TWOSIDES (Tatonetti et al., 2012)	DDI matrices $\mathbf{M}_{\text{DDI}}^{\text{bin}}, \mathbf{M}_{\text{DDI}}^{\text{freq}} \in \mathbb{R}^{ \mathcal{V}_{\text{med}} ^2}$	Safety verification (Section 4.4)
openFDA drug labels (Kass-Hout et al., 2016)	Contra. matrices $\mathbf{M}_{\text{Contra}}^{\text{bin}}, \mathbf{M}_{\text{Contra}}^{\text{freq}} \in \mathbb{R}^{ \mathcal{V}_{\text{med}} \times \mathcal{V}_{\text{diag}} }$	Safety verification (Section 4.4)

Table 7: Hierarchical taxonomy of diagnosis and medication concepts used in this study. Each level contains two attributes: *Name*, which specifies the semantic grouping at that level, and *# Concepts*, the number of unique codes at that level. Diagnosis categories are derived from ICD-CM codes mapped through the CCS diagnostic hierarchy, and medications follow the five-level ATC ontology. For ICD-CM levels, the *# Concepts* is reported as the number of ICD-9-CM and ICD-10-CM codes shown in parentheses, respectively.

Level	Description	Diagnosis	Medication
Level 1	Name # Concepts	Major category 18	Anatomical main group 14
Level 2	Name # Concepts	Subcategory 34	Therapeutic subgroup 90
Level 3	Name # Concepts	CCS code 265	Pharmacological subgroup 248
Level 4	Name # Concepts	Pre-decimal ICD-CM code (937, 1278)	Chemical subgroup 840
Level 5	Name # Concepts	Full ICD-CM code (13193, 73066)	Chemical substance 5495

records are organized into clinically meaningful fields, such as indications, contraindications, adverse reactions, and warnings, although field coverage and text format vary across products. The openFDA Drug Labeling data are updated over time as new safety and effectiveness information becomes available.

We use the openFDA Drug Labeling snapshot last updated on 2026-03-04. To construct the ATC-L4 contraindication knowledge database, we first merge all 13 openFDA Drug Labeling files from <https://open.fda.gov/data/downloads/>. We retain three fields: `substance_name`, `RxCUI`, and `contraindications`. We remove records without `RxCUI` (193,958 entries) and records without contraindications (27,769 entries). For duplicate single-`RxCUI` records, we group records by `RxCUI` and retain the record with the longest contraindication text, measured by `len(contraindications)`. This step reduces single-`RxCUI` records from 22,382 to 2,765 unique entries. We then remove all 11,152 multi-`RxCUI` records to avoid ambiguous drug-to-code mappings. Finally, we map the 2,765 single-`RxCUI` records to ATC-L4 codes and retain only successfully mapped

entries. This process yields 4,434 `RxCUI`-ATC-L4 pairs, corresponding to 2,136 unique `RxCUI` drugs and 400 unique ATC-L4 codes.

A.5 Mapping Coverage

All identifier mappings described above are deterministic and ontology-based: source drug identifiers are normalized through RxNorm concept identifiers into the closed ATC-L4 vocabulary, with no learned, fuzzy, or manually curated matching. Source entities that have no ATC-L4 target are dropped rather than approximated, and multiple source relations that map to the same ATC-L4 pair are merged by union for the binary matrices and summed for the weighted matrices (Appendices A.3 and A.4). Table 8 reports the resulting coverage.

Unmapped source entities are predominantly compounds outside the ATC hierarchy, such as combination products, contrast agents, and over-the-counter preparations without an assigned ATC code. Because the safety metrics treat relations absent from a resource as safe by default (Appendix E), this incomplete coverage makes the reported DDI and contraindication rates lower bounds on the true safety burden over $\mathcal{V}_{\text{med}}$, for

Table 8: **Deterministic mapping of external resources into the ATC-L4 vocabulary.** Covered counts ATC-L4 codes in the full 840-code ATC-L4 space (Table 7); the prediction vocabulary $\mathcal{V}_{\text{med}}$ is the cohort-specific subset of this space (Table 10).

Resource	Source	Mapped	Dropped	Rate	Covered
MEDI-2	3,031	2,182	849	72.0%	674
openFDA	2,765	2,136	629	77.3%	400
TWOSIDES	645	556	89	86.2%	399

every method we compare.

B Expert Construction

This appendix describes how we construct the expert set used by SafeRx-Agent. The expert taxonomy is derived in two stages. First, we serialize each patient case using the structured EHR-to-text template in Figure 3. We encode the serialized cases into patient-level text embeddings and sweep the number of clusters over these embeddings. Second, after selecting the best cluster count, we inspect the diagnosis composition, ICD-10 chapter distribution, and representative cases of each cluster to assign clinically interpretable expert names and chapter mappings. In addition to these cluster-derived experts, we introduce an always-on universal expert to cover ICU-common supportive-care medications that are not specific to a single diagnosis chapter.

EHR-to-text serialization. Each structured EHR case is serialized into natural language for LLM-based methods (Section 4 and the LLM/Agent baselines in Section 5). The serialized input contains the patient profile, historical visits (diagnoses, procedures, prescribed medications), and the target visit (diagnoses and procedures only). Each clinical concept is paired with its original code to preserve both textual semantics and code-level grounding. The full template is shown in Fig. 3.

B.1 Cluster-Derived Specialty Experts

We construct patient-level routing features from diagnosis information and apply K-means clustering with different values of K . Specifically, we sweep $K \in \{2, 3, \dots, 10\}$ and select the value with the best silhouette score. The sweep identifies $K = 7$ as the most separable solution, achieving a silhouette score of 0.45. This suggests that the patient population is best organized into seven coarse clinical groups under the constructed routing representation.

Figure 4 visualizes the resulting clusters. The

centroid–domain heatmap (Figure 4a) shows that each of the seven clusters concentrates its mass on a single clinical domain, with off-diagonal mass remaining low; this one-cluster-per-domain structure motivates a one-to-one cluster-to-expert assignment. The PCA projection of single-domain patients (Figure 4b) confirms that patients with one dominant clinical domain separate visibly along the first two principal components (53.9% + 18.5% explained variance), with obstetrics/perinatal forming a clearly distinct branch and the remaining six domains spread along the second axis.

We then examine each cluster’s dominant diagnosis chapters, diagnosis descriptions, and representative cases. Based on this inspection, we assign each cluster a clinically interpretable expert name and map it to the corresponding ICD-10 chapter group. Table 9 summarizes the resulting expert taxonomy.

Expert domain	ICD-10 chapters
Oncology/Hematology	II, III
Endocrine/Metabolic	IV
Cardiovascular	IX
Respiratory	X
Gastroenterology/Hepatology	XI
Musculoskeletal/Trauma	XII, XIII, XIX, XX
Obstetrics/Perinatal	XV, XVI, XVII

Table 9: Cluster-derived specialty experts and their ICD-10 chapter mappings. The expert names and chapter mappings are curated after selecting $K = 7$ from the clustering sweep.

Offline construction and inference-time determinism.

The expert taxonomy is constructed once, before evaluation, and is never re-derived per case: the clustering sweep over K selects the partition, the resulting clusters are mapped to seven ICD chapter-level specialties by manual inspection, and the panel is then frozen. Clustering therefore only defines the specialty taxonomy and is not part of inference. At inference, ROUTE is fully deterministic, using ICD chapter overlap and diagnosis keywords over the patient’s current and historical diagnoses together with the predefined expert scopes; it never accesses target or held-out medication labels. The closed output vocabulary $\mathcal{V}_{\text{med}}$ is likewise defined deterministically as the union of normalized ATC-L4 codes in the preprocessed cohort, with no frequency filtering and no evaluation-driven selection, and the same vocabulary and out-of-vocabulary handling are shared by every method we compare.

EHR-to-text serialization template

Patient profile. The patient’s age is [AGE] and gender is [GENDER]. The patient’s insurance type is [INSURANCE], language is [LANGUAGE], admission type is [ADMISSION TYPE], marital status is [MARITAL STATUS], and race is [RACE]. The patient has [T] ICU visits.

Historical visit. In visit [t], the patient had diagnoses: [DIAGNOSIS DESCRIPTION] ([ICD CODE]), ...; procedures: [PROCEDURE DESCRIPTION] ([PROCEDURE CODE]), ... The patient was prescribed drugs: [DRUG NAME] ([ATC-L4 CODE]), ...

Target visit. In this visit, the patient has diagnoses: [DIAGNOSIS DESCRIPTION] ([ICD CODE]), ...; procedures: [PROCEDURE DESCRIPTION] ([PROCEDURE CODE]), ... Then, the patient should be prescribed:

Figure 3: **Structured EHR-to-text serialization template.** Textual descriptions are paired with structured codes to preserve both clinical semantics and code-level grounding.

C Dataset Statistics

Table 10: Dataset statistics and splits after preprocessing. All splits are patient-level. For MIMIC-IV, 100 cases are held out for prompt development and the remaining 1,586 form the evaluation set used by all LLM-based methods; MIMIC-III is used as an external evaluation cohort with prompts transferred from MIMIC-IV. Supervised baselines require task-specific training and are trained, tuned, and evaluated on training, validation, and test splits of the corresponding cohort, containing 1,124, 281, and 281 patients for MIMIC-IV, so their results in Tables 1 and 2 are computed on the held-out test split rather than on the full evaluation set. Avg. Meds is the average ground-truth medication-set size, measured on the evaluation set; Prompt and Eval. are the prompt-development and evaluation set sizes.

Dataset	Cohort	$ V_{med} $	Avg. Meds	Prompt	Eval.
MIMIC-III	901	185	5.36	–	901
MIMIC-IV	1,686	274	9.17	100	1,586

D Implementation Details

This appendix provides additional implementation details for LLM inference, agent-side ablation settings, and prompt templates. All LLM-based methods are evaluated with the same serialized EHR input format, closed ATC-L4 medication vocabulary, and evaluation script unless otherwise specified.

D.1 Inference Settings

Table 11 summarizes the model-level inference settings used in our experiments. We generate one medication set per case. Closed-source models are accessed through their official APIs, while open-source models are served locally via vLLM. For backbone LLMs, we use temperature 0.0 for deterministic stages (summarization, critique, and safety verification) and 0.2 for the expert generation stage to encourage diverse medication proposals across specialists.

Table 11: **Inference settings for LLM-based methods.** Temp. column shows temperatures for deterministic / generation stages; medical LLM baselines use only direct prompting (single stage). Max model length denotes the maximum context length used for vLLM serving.

Model	Access	Temp.	Max Model Len.
<i>LLM backbones</i>			
GPT-5.2	API	0.0 / 0.2	–
Claude Sonnet-4.6	API	0.0 / 0.2	–
Qwen3-32B	Local (vLLM)	0.0 / 0.2	8,192
Gemma3-27B-IT	Local (vLLM)	0.0 / 0.2	8,192
<i>Medical LLM baselines</i>			
UltraMedical-70B	Local (vLLM)	0.0	4,096
Med42-v2-70B	Local (vLLM)	0.0	4,096
Llama3.1-Aloe-Beta-70B	Local (vLLM)	0.0	4,096
HuatuoGPT-o1	Local (vLLM)	0.0	4,096
OpenBioLLM-70B	Local (vLLM)	0.0	4,096

E Evaluation Metrics

We evaluate each prediction along two axes: *accuracy* and *safety*. Prediction accuracy is measured using Jaccard, precision, recall, and F1. Safety is measured using GT-normalized drug–drug interaction rates and drug–diagnosis contraindication rates. Let N denote the number of test visits; $\mathcal{M}^{(t)}$ and $\hat{\mathcal{M}}^{(t)}$ denote the ground-truth and predicted ATC-L4 medication sets for visit t ; and $\mathcal{D}^{(t)}$ denote the diagnosis set used for contraindication screening at that visit.

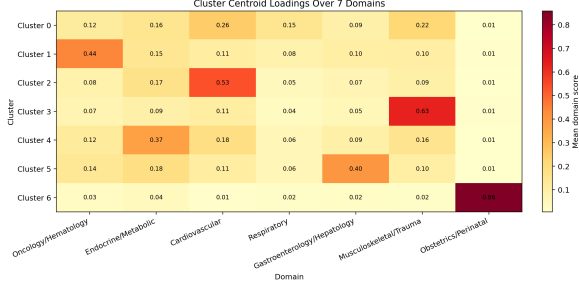
E.1 Prediction Accuracy Metrics

Jaccard. We report the sample-averaged Jaccard similarity:

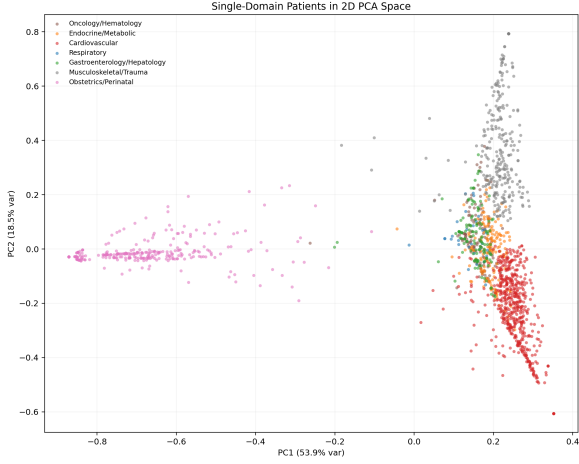
$$\text{Jaccard} = \frac{1}{N} \sum_{t=1}^N \frac{|\mathcal{M}^{(t)} \cap \hat{\mathcal{M}}^{(t)}|}{|\mathcal{M}^{(t)} \cup \hat{\mathcal{M}}^{(t)}|}. \quad (1)$$

Precision, Recall, and F1. We report micro-averaged precision, recall, and F1 by pooling true positives, false positives, and false negatives across all visits:

$$\text{Precision} = \frac{\sum_{t=1}^N |\mathcal{M}^{(t)} \cap \hat{\mathcal{M}}^{(t)}|}{\sum_{t=1}^N |\hat{\mathcal{M}}^{(t)}|}, \quad (2)$$



(a) Cluster centroid loadings over the seven derived clinical domains. Each cluster concentrates on a single domain.



(b) Single-domain patients projected onto the first two principal components of the routing feature space, colored by dominant clinical domain.

Figure 4: **Cluster-domain structure of the routing feature space.** The centroid heatmap (top) and PCA projection (bottom) jointly justify the $K = 7$ partition and the cluster-to-expert mapping in Table 9.

$$\text{Recall} = \frac{\sum_{t=1}^N |\mathcal{M}^{(t)} \cap \hat{\mathcal{M}}^{(t)}|}{\sum_{t=1}^N |\mathcal{M}^{(t)}|}, \quad (3)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

Micro averaging weights each drug-level decision equally and avoids giving the same weight to visits with very small and very large medication sets.

E.2 Prediction Safety Metrics

GT-normalized DDI rate. Drug-drug interactions are represented by a symmetric matrix $\mathbf{M}_{\text{DDI}} \in [0, 1]^{|\mathcal{V}_{\text{med}}| \times |\mathcal{V}_{\text{med}}|}$ indexed over the ATC-L4 medication vocabulary. We report two variants: DDI-B uses a binary matrix indicating whether any documented interaction exists, while DDI-W uses a weighted matrix with normalized interaction weights.

For each visit t , we first compute the predicted

interaction burden:

$$I_{\text{DDI}}^{(t)} = \sum_{i < j, i, j \in \mathcal{M}^{(t)}} \mathbf{M}_{\text{DDI}}[i, j]. \quad (5)$$

We normalize this burden by the number of possible ground-truth drug pairs:

$$Z_{\text{DDI}}^{(t)} = \max\left(\frac{I_{\text{DDI}}^{(t)}}{\binom{|\mathcal{M}^{(t)}|}{2}}, 1\right). \quad (6)$$

The per-visit GT-normalized DDI score is:

$$s_{\text{DDI}}^{(t)} = \min\left(\frac{I_{\text{DDI}}^{(t)}}{Z_{\text{DDI}}^{(t)}}, 1\right). \quad (7)$$

We report the dataset-level DDI score as:

$$\text{DDI}_{\text{GTnorm}} = \frac{1}{N} \sum_{t=1}^N s_{\text{DDI}}^{(t)}. \quad (8)$$

Unlike the conventional DDI rate, which normalizes by the number of predicted drug pairs, this metric normalizes the predicted interaction burden by the expected prescription complexity of the corresponding visit. This prevents large prediction sets from artificially lowering the DDI score by adding many non-interacting pairs. The per-visit score is capped at 1 so that visits with excessive interaction burden are treated as maximally unsafe without dominating the dataset average. Lower values indicate safer predictions.

GT-normalized contraindication rate. Drug-diagnosis contraindications are represented by a bipartite matrix $\mathbf{M}_{\text{Contra}} \in [0, 1]^{|\mathcal{V}_{\text{med}}| \times |\mathcal{V}_{\text{diag}}|}$, where $\mathbf{M}_{\text{Contra}}[m, d]$ indicates whether medication m is contraindicated for diagnosis d under the binary variant, or gives a normalized contraindication weight under the weighted variant.

For each visit t , we first compute the predicted contraindication burden:

$$I_{\text{contra}}^{(t)} = \sum_{m \in \mathcal{M}^{(t)}} \sum_{d \in \mathcal{D}^{(t)}} \mathbf{M}_{\text{Contra}}[m, d]. \quad (9)$$

We normalize this burden by the ground-truth medication-set size and the number of active diagnoses:

$$Z_{\text{contra}}^{(t)} = \max\left(\frac{I_{\text{contra}}^{(t)}}{|\mathcal{M}^{(t)}| |\mathcal{D}^{(t)}|}, 1\right). \quad (10)$$

The per-visit GT-normalized contraindication score is:

$$s_{\text{contra}}^{(t)} = \min\left(\frac{I_{\text{contra}}^{(t)}}{Z_{\text{contra}}^{(t)}}, 1\right). \quad (11)$$

Table 12: **Case-level paired bootstrap** (5,000 resamples, Gemma3-27B-IT). GA = General Agent. Positive Δ favors SafeRx-Agent for F1; negative Δ indicates reduced risk for DDI-W. † Measured before safety verification for both systems, matching Table 4.

Dataset	Comparison	Δ	95% CI	p
<i>Prediction quality (F1)</i>				
MIMIC-IV	vs. RareAgents	+0.039	[+0.009, +0.093]	0.019
MIMIC-III	vs. GA	+0.116	[+0.077, +0.189]	< 0.005
MIMIC-IV	vs. call-matched GA †	+0.053	[+0.018, +0.103]	0.006
<i>Safety effect of the verifier (DDI-W)</i>				
MIMIC-IV	w/ vs. w/o verifier	-0.290	[-0.491, -0.137]	< 0.001
MIMIC-III	w/ vs. w/o verifier	-0.401	[-0.569, -0.140]	< 0.001

We report the dataset-level contraindication score as:

$$\text{Contra}_{\text{GTnorm}} = \frac{1}{N} \sum_{t=1}^N s_{\text{contra}}^{(t)}. \quad (12)$$

Indices that fall outside the matrix support contribute zero to the numerator under both metrics: drug pairs (i, j) outside $\mathbf{M}_{\text{DDI}}$ for DDI-B / DDI-W and drug–diagnosis pairs (m, d) outside $\mathbf{M}_{\text{Contra}}$ for Contra-B / Contra-W. This convention treats safety-knowledge-unknown pairs as safe-by-default, so the reported DDI / contraindication rates are lower bounds on the true safety burden over $\mathcal{V}_{\text{med}}$. Lower values indicate safer predictions.

F Statistical Significance and Run-to-Run Stability

This appendix complements Tables 1 and 2, which report a single run per configuration, with inferential and stability analyses on the primary analysis setting (Gemma3-27B-IT), over the complete MIMIC-IV (1,586-case) and MIMIC-III (901-case) evaluation cohorts. Unless marked otherwise, rows use the full SafeRx-Agent pipeline with safety verification, that is, the *w/ safety filter* rows of Tables 1 and 2.

Paired bootstrap. Table 12 reports case-level paired bootstrap tests with 5,000 resamples. Resampling is paired at the patient level, so each resample scores both systems on the same cases and the interval reflects only disagreement between methods rather than cohort variation. Both the prediction margin over the strongest agent baseline on each dataset and the safety effect of the verifier are significant, and no confidence interval crosses zero.

Run-to-run stability. Because SafeRx-Agent samples at the expert generation stage (Appendix D.1), we additionally re-run the full pipeline end to end with three sampling seeds (Table 13).

Table 13: **Run-to-run stability** of the full SafeRx-Agent pipeline over three sampling seeds (Gemma3-27B-IT).

Dataset	Metric	Seed 1	Seed 2	Seed 3	Mean $\pm$ Std
MIMIC-IV	F1	0.3718	0.3809	0.3757	0.3761 $\pm$ 0.0045
	DDI-B	0.3456	0.3629	0.3906	0.3664 $\pm$ 0.0185
MIMIC-III	F1	0.2620	0.2652	0.2625	0.2632 $\pm$ 0.0014
	DDI-B	0.3810	0.3652	0.3932	0.3798 $\pm$ 0.0115

Table 14: **Verifier decision audit** over all flagged items (MIMIC-IV, Gemma3-27B-IT).

Audit item	Rate
Retention with documented prior exposure	83.5%
Retention without prior exposure	64.7%
Removal under an active contraindication	97.9%

Across seeds, F1 varies by at most 0.005 and DDI-B by at most 0.019, an order of magnitude below the margins in Table 12, so the reported gains are not attributable to sampling variation. The same qualitative patterns hold for the other three backbones in Tables 1 and 2.

G Verifier Behavior Audit

Section 4.4 specifies VERIFY as a constrained adjudicator: FINDFLAGS retrieves every candidate risk deterministically from $\mathbf{M}_{\text{DDI}}$ and $\mathbf{M}_{\text{Contra}}$, and VERIFY chooses an action in $\{\text{RET}, \text{REM}\}$ for each flag, so it can neither introduce a medication outside the candidate set nor assert an undocumented safety relation. This appendix checks whether its decisions follow the intended patient-conditioned policy, auditing all flagged items in the MIMIC-IV evaluation cohort (Table 14).

The verifier retains flagged medications far more often when the patient was already exposed to them (83.5% versus 64.7%, an 18.8-point gap), and almost always removes a medication contraindicated against an active diagnosis (97.9%). Its behavior is therefore neither indiscriminate deletion of every flagged medication nor blanket retention: it separates continuation-supported risks from unsupported ones, as the adjudication policy in Figure 10 intends.

This audit tests policy consistency rather than clinical correctness. Verifying that individual retain/remove rulings are clinically appropriate would require adjudication by pharmacists or physicians, which we leave to future work.

Transfer to a matched generation stack. The verifier is independent of how candidates were produced, so we can apply the identical module to

Table 15: **The identical safety verifier applied to two generation stacks** (MIMIC-IV, Gemma3-27B-IT). All metrics are lower-is-better; pre and post denote before and after verification.

Generation stack	DDI-B		DDI-W		Contra-B	
	pre	post	pre	post	pre	post
Call-matched GA	0.7196	0.5970	0.5414	0.4250	0.0091	0.0058
SafeRx-Agent	0.6683	0.3695	0.5761	0.2865	0.0086	0.0030

Table 16: **Set-size-matched deletion controls** (MIMIC-IV, Gemma3-27B-IT). Each policy deletes the same per-case number of medications from the pre-verification set as VERIFY removes, so every matched row has identical per-case set size. All safety metrics are lower-is-better.

Deletion policy	#Pred	DDI-B	DDI-W	Contra-B
No deletion (pre-verification)	9.65	0.6683	0.5761	0.0086
<i>Patient-agnostic deletion rules</i>				
Random	6.79	0.6640	0.4105	0.0046
Lowest generator confidence	6.79	0.6631	0.4053	0.0046
Lowest global frequency	6.79	0.6223	0.4169	0.0049
Highest global DDI degree	6.79	0.5479	0.3118	0.0044
<i>Patient-conditioned adjudication</i>				
SafeRx VERIFY	6.79	0.3695	0.2865	0.0030

a different generation stack. Applying it to the call-matched general agent of Table 4 reduces documented risk there as well (Table 15), which shows that the safety gains reported in Section 6.2 come from the verification step rather than from any interaction with SafeRx-Agent’s specialty decomposition. SafeRx-Agent nevertheless remains substantially safer after verification, so the overall advantage in Tables 1 and 2 is not attributable to the verifier alone.

Set-size-matched deletion controls. Table 16 reports the size-matched controls of Section 6.4 in full. Each policy deletes exactly as many medications per case as VERIFY removes, so every matched row has the same per-case set size and the rows differ only in *which* medications are dropped. The four control rules ignore the patient: random deletion, deletion by lowest generator confidence, deletion of the globally least frequent ATC-L4 classes, and deletion of the drugs with the largest global DDI degree. Only VERIFY conditions each decision on the individual record. Deleting by global DDI degree is the strongest control, as expected, since it directly targets the quantity DDI-B measures, yet VERIFY still reduces DDI-B by a further 0.178 (0.5479 to 0.3695) and Contra-B from 0.0044 to 0.0030. The safety gain is therefore risk-targeted rather than a consequence of predicting fewer medications.

Table 17: **Critique action distribution.** Share% is the fraction of all critique decisions assigned to each action. TP% and FP% report the composition of retained and removed medications with respect to the ground-truth medication set.

Action	Count	Share%	TP%	FP%
RET	16098	70.8	37.8	62.2
REM	6625	29.2	11.7	88.3

Table 18: **Per-expert removal behavior.** Rm% denotes the fraction of proposed medications removed by critique. RmFP% denotes the false-positive rate among removed medications.

Expert	Proposed	Removed	Rm%	RmFP%
Universal supportive	12345	822	6.7	84.5
Endocrine/metabolic	4194	745	17.8	85.4
Cardiovascular	4847	1128	23.3	82.7
Gastroenterology/hepatology	2391	602	25.2	84.6
Musculoskeletal/trauma	4907	1420	28.9	88.1
Oncology/hematology	2993	920	30.7	93.2
Obstetrics/perinatal	2344	758	32.3	96.6
Respiratory	1308	490	37.5	90.6

H Additional Diagnostic Analysis

Figure 5 shows the diagnostic and expert analyses discussed in Sections 6.5 and 6.6. This appendix then complements them with two additional diagnostics for the critique stage: the overall action distribution (Table 17) and per-expert removal behavior (Table 18).

Removal behavior differs across experts. The universal supportive-care expert has the lowest removal rate, consistent with its role in proposing common ICU medications. Specialty experts have higher removal rates and consistently high RmFP%, indicating that critique mostly filters overly broad or weakly supported specialty-specific proposals rather than indiscriminately pruning correct ones.

I Additional Expert Analysis

This appendix complements the leave-one-expert-out analysis in Section 6.6 with two further diagnostics: router activation behavior (Figure 6) and expert-level proposal/retention statistics (Table 19). Expert abbreviations used throughout this appendix and in Figure 5c: SUP = universal supportive, CVD = cardiovascular, ENDO = endocrine/metabolic, MSK = musculoskeletal/trauma, ONC = oncology/hematology, OB = obstetrics/perinatal, GI = gastroenterology/hepatology, RESP = respiratory.

The cardiovascular expert has the highest TP/Ret. ratio (51.5%), suggesting that specialty experts contribute targeted domain-specific candidates beyond what broad supportive-care coverage provides; spe-

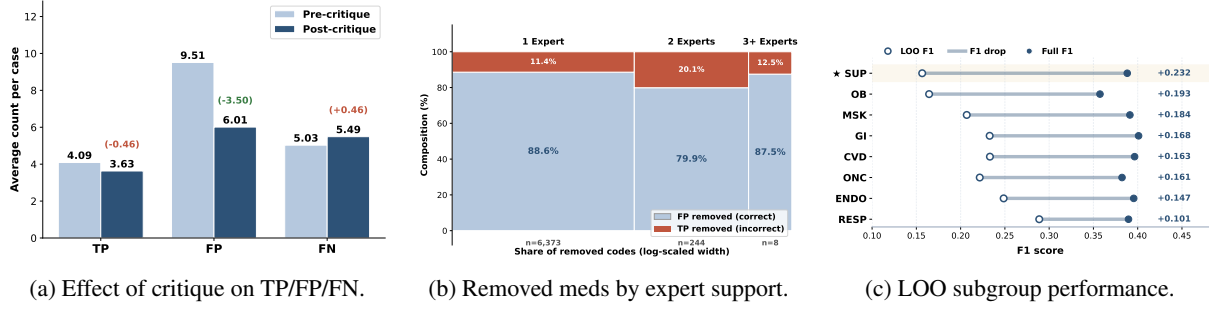


Figure 5: **Diagnostic and expert analyses.** (a) Critique substantially reduces average false positives. (b) Critique primarily removes medications proposed by only one expert, and effectively removes false-positives. (c) Leave-one-expert-out subgroup performance; expert abbreviations are defined at the beginning of Appendix I.

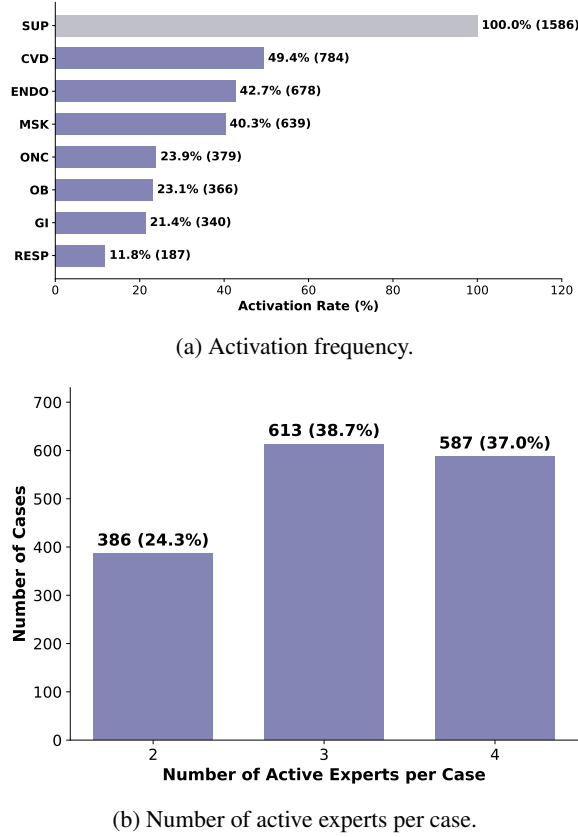


Figure 6: **Expert activation statistics.** The router activates a sparse, case-dependent subset of experts. Expert abbreviations are defined at the beginning of Appendix I.

cialty experts that are activated less frequently (e.g., RESP, OB) still maintain TP/Ret. above 35%, indicating that the router does not over-activate marginal experts.

J Routing Coverage

ROUTE (Section 4.3) is a multi-label rule rather than a single-best selector: it matches all current and historical ICD codes in X_T against every expert scope, so several specialty experts can be activated at once and the always-on supportive-care

expert is never omitted. There is consequently no unique gold specialty set for a case, and a conventional routing error rate is undefined. The substantive concern is instead whether the non-routed experts would have recovered clinically relevant medications that sparse routing misses. We test this directly by comparing sparse routing against full-panel activation (Table 20).

Activating every expert raises recall by only 0.030 while lowering precision by 0.039 and F1 by 0.009, raising DDI-B by 0.2015 and Contra-B by 0.0012, and doubling the number of generation-stage LLM calls (17.0 versus 8.4). Indiscriminate activation therefore mainly injects off-scope candidates that the routed subset already excludes, and the few medications recovered by non-routed experts are outweighed by the added false positives and safety violations. Diagnosis-based multi-label routing thus gives multi-comorbid patients broad coverage without the precision and safety cost of full activation. The rule depends only on the patient’s current and historical diagnoses and the pre-defined expert scopes; it never accesses target or held-out medication labels.

K Efficiency Breakdown

Table 21 reports the average LLM calls, input tokens, output tokens, and wall-clock latency per case for each operator in Algorithm 1. SUMMARIZE, GENERATE, and VERIFY are invoked per activated expert or per flagged item, so their per-case totals scale with case complexity. ROUTE (Section 4.3) and FINDFLAGS (Section 4.4) are deterministic and contribute negligible cost.

Call-matched general agent. The call-matched control in Table 4 repeats the general agent’s generation and critique stack over independent self-consistency passes and aggregates the results, con-

Table 19: **Expert contribution statistics.** Activated counts selected cases. Proposed and retained count expert-level medication candidates before and after critic filtering. TP/Ret. is the true-positive density among retained predictions.

Expert	Activated	Proposed	Retained	Ret.%	TP	TP/Ret.
Universal supportive	1586	12345	11523	93.3	4573	39.7
Cardiovascular	832	4847	3719	76.7	1915	51.5
Musculoskeletal/trauma	672	4907	3487	71.1	1633	46.8
Endocrine/metabolic	721	4194	3449	82.2	1581	45.8
Gastroenterology/hepatology	360	2391	1789	74.8	796	44.5
Oncology/hematology	405	2993	2073	69.3	874	42.2
Respiratory	197	1308	818	62.5	354	43.3
Obstetrics/perinatal	393	2344	1586	67.7	570	35.9

Table 20: **Sparse routing versus full-panel activation** (MIMIC-IV, Gemma3-27B-IT, before safety verification). Calls counts generation-stack LLM calls per case: two per activated expert plus one critique.

Activation	Exp.	Calls	Prec.	Rec.	F1	DDI-B	Contra-B
ROUTE	3.7	8.4	0.379	0.398	0.388	0.6683	0.0086
All experts	8.0	17.0	0.340	0.428	0.379	0.8698	0.0098

Table 21: **Per-phase efficiency breakdown of SafeRx-Agent on MIMIC-IV.** All numbers are averaged over evaluation cases. ROUTE and FINDFLAGS are deterministic (no LLM calls).

Phase	Calls	In Tokens	Out Tokens	Sec.
<i>Multi-agent generation</i>				
ROUTE	0	0	0	<0.01
SUMMARIZE	3.7	5,010	2,559	5.7
GENERATE	3.7	11,475	2,329	13.1
CRITIQUE	1.0	7,775	1,100	8.9
<i>Safety verification</i>				
FINDFLAGS	0	0	0	<0.01
VERIFY	1.0	587	82	1.8
Total w/o safety filter	8.4	24,260	5,988	27.7
Total w/ safety filter	9.4	24,847	6,070	29.5

suming 10.0 LLM calls per case on average, above SafeRx-Agent’s 8.4 generation calls and 9.4 total calls. We match model calls rather than input tokens because SafeRx-Agent distributes short, specialty-scoped contexts across experts while each generalist pass consumes the full serialized record, so equalizing token counts would require truncating the generalist’s context or padding SafeRx-Agent’s, either of which changes the method being measured.

L Operating Points of Alternative Approaches

Section 6.8 compares SafeRx-Agent against the LLM and agent baselines under identical serving conditions. Table 22 places these methods, together with lightweight supervised models, on the axes that matter for deployment. The comparison is qualitative for the supervised row: those models run a single local forward pass rather than issuing

Table 22: **Deployment operating points.** Online cost is taken from Table 5 where measured. Trace = granularity of the intermediate decisions exposed to a reviewer.

Approach	Training	Online cost	Evidence	Trace
Supervised DL	Per dataset	1 fwd. pass	–	–
Direct prompting	None	1.0 call, 9.6s	–	–
General agent	None	3.0 calls, 10.2s	✓	Single-agent
RareAgents	None	10.1 calls, 13.3s	✓	Tool-level
SafeRx-Agent	None	9.4 calls, 29.5s	✓	Per-expert

API calls, so their wall-clock latency is not comparable to the served LLM rows and we do not report a number for it.

Supervised models are preferable when minimum serving latency is the primary objective. They require task- and dataset-specific training, however, and produce fixed-function predictions without inference-time evidence retrieval, traceable intermediate decisions, or case-specific adjudication of drug interactions and contraindications; making them safety-aware typically requires encoding predefined safety objectives into the training loss. SafeRx-Agent is training-free at deployment, runs on multiple backbones, and transfers from MIMIC-IV to MIMIC-III without prompt re-tuning. It therefore does not claim the lowest achievable latency: it spends additional inference compute to obtain evidence grounding, per-medication traceability, and explicit safety verification.

M Prompt Templates

This section provides all prompt templates used in our experiments. Appendix M.1 lists the templates for the LLM-based operators in SafeRx-Agent, and Appendix M.2 lists the baseline templates for direct prompting, the general-agent baseline, and the RareAgents-style baseline.

M.1 SafeRx-Agent Operator Prompts

Figures 7, 8, 9, and 10 correspond to SUMMARIZE, GENERATE, CRITIQUE, and VERIFY, respectively. The non-LLM operators ROUTE and FINDFLAGS

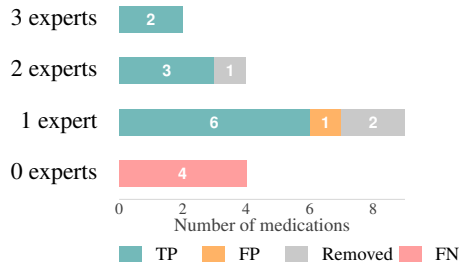


Figure 15: **Prediction outcome by expert support** (case study, Appendix N). Medications proposed by more experts are more likely to be true positives. All 5 codes with ≥ 2 expert support that were retained by CRITIQUE are correct (TP). The only false positive (A06AB) and all 3 correctly removed codes were single-expert proposals. The 4 FN medications were not proposed by any expert.

are deterministic and do not require prompt templates.

M.2 Baseline Prompt Templates

We provide abbreviated prompt templates used in our experiments. Figures 11–13 show the baseline templates for direct prompting, the general-agent baseline, and the adapted RareAgents baseline.

N Case Study

We present an end-to-end walkthrough of SafeRx-Agent on a representative MIMIC-IV case (Subject 10785159) using Gemma3-27B-IT. This case illustrates how multi-expert routing, grounded generation, and global critique produce a traceable medication set with per-medication rationales. Figure 14 provides an overview of the pipeline flow on this case; Table 23 traces every candidate code through the system.

N.1 Patient Presentation



Patient Vignette

An 87-year-old female with a history of nonrheumatic aortic stenosis (post-bioprosthetic valve replacement), chronic heart failure, persistent atrial fibrillation, cardiomyopathy, hypertensive heart disease, type 2 diabetes mellitus on insulin, hyperlipidemia, obesity, GERD, irritable bowel syndrome, and prior venous thromboembolism. She presents with **NSTEMI**, acute-on-chronic systolic heart failure, COVID-19, pulmonary hypertension, phlebitis of the lower extremities, and hypotension. Current visit procedures include coronary angiography and cardiac catheterization. Prior visit medications (12 codes) include amiodarone (C01BD), furosemide (C03CA), metoprolol (C07AB), heparin (B01AB), PPI (A02BC), potassium (A12BA), magnesium (A12CC), tramadol (N02AX), contact laxatives (A06AB), osmotic laxatives (A06AD), melatonin (N05CH), and carbohydrate supplement (V06DC).

Ground-truth medications for the current visit: 15 ATC-L4 codes.

N.2 Pipeline Walkthrough

Step 1: Expert Routing (ROUTE). ROUTE scores the patient’s ICD codes against each expert’s scope. The cardiovascular domain dominates (score = 0.60), followed by endocrine/metabolic (0.17). Three experts are activated: *Cardiovascular*, *Endocrine/Metabolic*, and the always-on *Universal Supportive Care* (Figure 14, rows 2–3).

Step 2: Expert Summarization (SUMMARIZE). Each activated expert extracts specialty-relevant evidence from the patient record:

- **Cardiovascular** identifies NSTEMI, acute systolic HF, AF, aortic stenosis, pulmonary HTN, and hypotension as active problems. It notes the prior aortic valve replacement and long-term anticoagulant use, and flags risks of bleeding (anticoagulation plus recent procedure), arrhythmia (AF, post-MI), and cardiorenal syndrome.
- **Endocrine/Metabolic** focuses on type 2 diabetes with long-term insulin use, and flags electrolyte management (prior potassium and magnesium supplementation) and metabolic instability risks given acute heart failure.
- **Universal Supportive Care** provides a broad view covering pain management (NSTEMI), bowel care (IBS history plus opioid use), sleep support, GI prophylaxis (GERD plus anticoagulation), and electrolyte repletion in the context of heart failure and anticipated diuretic use.

Step 3: Per-Expert Generation (GENERATE). Each expert proposes ATC-L4 codes grounded in its summary and the MEDI indication resource. The Cardiovascular expert proposes 7 codes, the Endocrine/Metabolic expert proposes 5 codes, and the Universal Supportive Care expert proposes 11 codes, yielding 15 unique candidates across all experts (Table 23).

Step 4: Global Critique (CRITIQUE). CRITIQUE reviews all expert proposals under the full patient context and removes three candidates:

- **C09AA** (ACE inhibitors): while plausible given diabetes and cardiovascular disease, the acute presentation (hypotension, acute HF) does not strongly support initiation, and no prior ACE inhibitor use is documented.
- **C10AA** (statins): both proposing experts assigned low confidence, and the acute presentation does not prioritize statin initiation.
- **A04AA** (antiemetics): a speculative proposal, as COVID-19 alone does not necessitate antiemetics without documented nausea or vomiting.

The final candidate set contains 12 ATC-L4 codes.

N.3 Analysis

The final predicted set contains 12 medications, achieving a precision of 0.917, recall of 0.733, and F1 of 0.815. Figure 15 summarizes the relationship between expert support and prediction outcome.

Multi-expert corroboration. All five codes proposed by ≥ 2 experts that were retained by CRITIQUE are true positives (C07AB, A12BA, B01AB, C01BD, A02BC), consistent with the aggregate finding that multi-expert corroboration is a strong signal for correctness (Section 6.5).

Selective critique. CRITIQUE removes 3 of the 15 candidate codes. None of the three removed drugs appear in the ground truth, indicating that the CRITIQUE correctly identified false positives in this case. Among single-expert proposals, CRITIQUE retains those with clear patient-specific support—for example, C03CA (loop diuretics) was proposed only by the Cardiovascular expert but was kept because the patient has documented heart failure, volume overload indicators, and prior furosemide use. In contrast, C09AA (ACE inhibitors) was removed despite cardiovascular relevance because

the patient's hypotension and lack of prior ACE inhibitor use argued against initiation.

False positive. The single false positive, A06AB (contact laxatives), was prescribed in the prior visit and proposed by the Endocrine/Metabolic expert based on documented constipation history. This is a clinically reasonable continuation that falls outside the current visit's ground-truth set, illustrating that not all false positives represent clinical errors.

False negatives. The four missed medications—A06AX (other drugs for constipation), D01AC (imidazole/triazole antifungals), J07BX (other viral vaccines), and N06AX (other antidepressants)—were not proposed by any expert. D01AC and J07BX lack explicit diagnostic indicators in the structured record (antifungal prophylaxis and COVID-19 vaccination are context-dependent decisions). N06AX (antidepressants) may reflect undocumented psychiatric history. These false negatives highlight the challenge of predicting medications whose indications require implicit clinical reasoning beyond the available EHR data.

Table 23: **Case study pipeline trace.** Checkmarks indicate which expert(s) proposed each ATC-L4 code via GENERATE. CRITIQUE action shows whether the code was retained (RET) or removed (REM). Outcome classifies each code against the ground-truth medication set: true positive (TP), false positive (FP), or not applicable (—) for correctly removed candidates. False-negative medications (bottom) were not proposed by any expert.

ATC-L4	Drug Class	CV	Endo	Univ	#Exp	Critique	Outcome
C07AB	Beta-blockers, selective	✓	✓	✓	3	RET	TP
A12BA	Potassium	✓	✓	✓	3	RET	TP
B01AB	Heparin group	✓		✓	2	RET	TP
C01BD	Antiarrhythmics, class III	✓		✓	2	RET	TP
A02BC	Proton pump inhibitors		✓	✓	2	RET	TP
C10AA	HMG CoA reductase inhibitors	✓	✓		2	REM	—
C03CA	Sulfonamides, plain (loop diuretics)	✓			1	RET	TP
C09AA	ACE inhibitors, plain	✓			1	REM	—
A06AB	Contact laxatives		✓		1	RET	FP
N02AX	Other opioids			✓	1	RET	TP
A06AD	Osmotically acting laxatives			✓	1	RET	TP
A12CC	Magnesium			✓	1	RET	TP
V06DC	Carbohydrates			✓	1	RET	TP
N05CH	Melatonin receptor agonists			✓	1	RET	TP
A04AA	Serotonin (5-HT ₃) antagonists			✓	1	REM	—
<i>Ground-truth medications not proposed by any expert (false negatives):</i>							
A06AX	Other drugs for constipation				0	—	FN
D01AC	Imidazole/triazole antifungals				0	—	FN
J07BX	Other viral vaccines				0	—	FN
N06AX	Other antidepressants				0	—	FN

SafeRx-Agent SUMMARIZE Prompt Constructor

System prompt:

You are the activated specialty module inside a routed inpatient medication prediction system. The upstream router selected this skill; use the routing selection as an interpretive lens and not as permission to invent facts not present in the patient record.

Input:

- Skill name and profile.
- Available tools.
- Summarization playbook (expert-specific).
- Patient record.

Task rules:

- Produce a specialty-aware medication summary by reading the patient record through the lens of the activated expert's playbook.
- Keep the summary grounded in the record only; do not invent diagnoses, procedures, or medications that are not documented.
- Highlight domain-specific risks, procedures, care context, and prior medication evidence that could affect inpatient medication classes.
- Do not predict medications at this stage; medication proposals are produced downstream by the GENERATE operator.

Output format:

Return strict JSON with five fields: *expertise* (the activated expert identity), *current_admission* (specialty-relevant active problems, organ dysfunction, and care context), *medication_relevant_history* (prior indications, procedures/devices, and explicit medication evidence), *expertise_focus* (record-grounded facts explaining why this expert is relevant), and *risks_to_watch* (specialty-relevant medication risks).

Example playbook (universal supportive expert):

Focus the summary on factors that drive universal supportive medication needs:

- **ICU / critical-care context:** ICU admission, intubation, vasopressor use, post-surgical, or otherwise critically ill – drives prophylactic needs.
- **Immobility / VTE risk:** bed-bound, post-operative, trauma, or prolonged hospitalization → VTE prophylaxis.
- **GI stress / aspiration risk:** NPO status, intubation, critical illness, steroid use → stress-ulcer prophylaxis, bowel management, antiemetics.
- **Pain / sedation needs:** surgical procedures, trauma, burns, invasive lines → analgesics, sedatives.
- **Electrolyte context:** renal function, diuretic use, critical illness → electrolyte monitoring and repletion.
- **Nutritional status:** NPO, TPN, tube feeding, prolonged admission → vitamins, IV fluids, dextrose.
- **Infection / vaccine status:** any hospitalization → influenza-vaccine consideration; critical illness → infection prophylaxis.
- **Prior medications:** note supportive medications from prior visits that suggest continuation.

Do not summarize specialty-specific clinical details (e.g., cardiac rhythm, tumor staging); those belong to routed specialty experts.

Figure 7: Prompt template for the SafeRx-Agent SUMMARIZE operator described in Section 4.3. The prompt is dynamically constructed per activated expert by injecting the expert-specific playbook into the shared template; the universal supportive expert's playbook is shown as an example.

SafeRx-Agent GENERATE Prompt Template

System prompt:

You are the medication generator for the activated expertise skill. This is a medication-list prediction task for the current admission; work only from the expert-specific summary and the patient record. Do not output guideline-only wish lists – predict what is plausibly on the medication list.

Input:

Expert-specific clinical summary (the JSON produced by SUMMARIZE); expert-specific drug-prediction checklist; prior-visit medication evidence; retrieved indication candidates from MEDI; and optional revision feedback from a prior generation round.

Task rules:

- Predict ATC-L4 medication classes likely present during the current admission. Output only 5-character ATC-L4 codes from the closed vocabulary; convert any ingredient- or product-level ATC codes to their ATC-L4 parent.
- Map from clinical scene to plausible medication class, then choose the fitting ATC-L4 code; do not return an empty list when the summary clearly supports inpatient medication classes.
- Follow this evidence priority (strongest first): (i) explicit medication evidence, (ii) prior-visit medications, (iii) procedure, device, or care-context evidence, (iv) current diagnoses and acute organ dysfunction, (v) retrieved indication candidates, (vi) supportive inpatient priors when supported by the documented scene.
- When a revision block is present, address the prior critique by adjusting only the affected codes; do not regenerate the entire list from scratch.

Output format:

Return strict JSON containing the predicted ATC-L4 codes, per-code confidence scores, input-grounded reasons that cite specific summary fields, any assumptions that limit certainty, and plausible alternative codes when the evidence is ambiguous.

Expert-specific checklist (example: universal supportive expert). The checklist gives per-class predict/withhold rules grounded in documented clinical conditions. For the universal supportive expert it covers four groups:

- **Thromboprophylaxis** (heparins, antiplatelets) — predict when VTE prophylaxis is documented, the patient is critically ill or post-operative, or antiplatelet therapy is indicated by coronary/cerebrovascular disease; withhold when active bleeding, HIT, or an anticoagulation contraindication is present.
- **Analgesia** (opioids, acetaminophen, salicylates) — predict for major surgery, severe trauma, documented pain or fever, or an active analgesia-sedation protocol; do not predict from mild pain, minor procedures, or ICU admission alone.
- **GI prophylaxis** (acid suppression, antiemetics, laxatives) — predict when GI bleeding history, stress-ulcer risk factors (mechanical ventilation *with* coagulopathy or shock), opioid-induced constipation, or postoperative nausea is documented; do not predict from ICU stay alone or in the presence of diarrhea or bowel obstruction.
- **Electrolyte and nutrition support** (calcium, potassium, dextrose, parenteral nutrition) — predict when specific electrolyte depletion, diuretic use, renal impairment, insulin infusion, or enteral/parenteral feeding is documented; do not predict from IV access or NPO status alone.

Figure 8: Prompt template for the SafeRx-Agent GENERATE operator described in Section 4.3. An expert-specific drug-prediction checklist is injected at runtime to constrain drug-class prediction to the activated expert's scope.

SafeRx-Agent CRITIQUE Prompt

System prompt:

You are the attending physician critique. Multiple specialist experts have proposed ATC-L4 medication classes for this inpatient admission. Your job is to produce the final prescription list by conservatively filtering the union of proposed codes.

Input:

Clinical summary; raw patient record; expert proposals with rationales; prior-visit medication evidence; and the union of proposed ATC-L4 codes.

Audit rules:

- Review each proposed code using the clinical summary, raw record, expert rationales, prior-medication evidence, and inpatient care context.
- Keep a code if it has concrete patient-specific support from any expert or record evidence; if it is a prior-visit medication whose underlying condition appears ongoing; or if it is a plausible inpatient supportive or prophylactic medication.
- Remove a code only if it is unsupported by any evidence in the summary, raw record, expert proposals, or prior-medication context, and has no plausible inpatient reason given the documented conditions, procedures, or care context.
- Do not remove a code simply because it is outside a specialist's scope. Do not remove evidence-grounded prophylactic or supportive medications unless there is specific evidence that they are contraindicated or absent.
- Remove redundant or overlapping codes only when they clearly duplicate the same therapeutic intent without added benefit.
- When in doubt, keep the code. Do not add codes outside the union.

Output format:

Return strict JSON containing the final retained ATC-L4 codes, removed codes with brief removal reasons, an overall rationale, and any missing information needed for clarification.

Figure 9: Abbreviated prompt template for the SafeRx-Agent CRITIQUE operator.

SafeRx-Agent VERIFY Prompt

System prompt:

You are a clinical pharmacist safety reviewer. Given a list of predicted prescription drugs and safety signals (drug–drug interactions and contraindications), decide which drugs to REMOVE or RETAIN.

Adjudication rules:

- If a drug is in `prior_meds` (the patient was on it before this admission), prefer to KEEP it unless the interaction is severe.
- When a DDI pair is flagged, prefer removing the drug with the higher global DDI degree (more interactions overall) and lower clinical necessity.
- When a drug is contraindicated with a patient diagnosis, remove it unless it was in `prior_meds` and the benefit clearly outweighs the risk.
- Be conservative: remove drugs only when there is a clear safety signal. Do not remove a drug merely because it has a low DDI degree.
- Return strict JSON only, with no markdown fences and no commentary.

Input (dynamically assembled user prompt):

The user prompt contains four sections, assembled at runtime from the candidate set and the two safety matrices:

1. **Predicted drugs.** Each candidate listed as `<ATC-L4>: <class name>`, with `[PRIOR-MED]` when the drug was active in the patient's previous admission. Example: `B01AB: Heparin group [PRIOR-MED]`.
2. **Prior medications.** All drugs active in the previous visit, used for continuation decisions.
3. **DDI pairs detected.** Each flagged pair from the binary DDI matrix, annotated with both drugs' global DDI degree and a `[PRIOR]` tag where applicable. Example: `B01AB (degree=42) [PRIOR] ↔ N02BA (degree=38)`.
4. **Contraindication pairs detected.** Each flagged drug–diagnosis pair from the binary contraindication matrix. Example: `A02BB ↔ diagnosis 080`.

If a case raises no DDI or contraindication flags, the candidate set is retained without invoking the verifier.

Output format:

Return strict JSON with two fields: `kept_drugs` (list of retained ATC-L4 codes) and `removed_drugs` (list of objects, each with a code and a short reason). The schema admits no other action: the verifier can only retain or remove members of the candidate set, and never emits a substitute or newly generated medication.

Figure 10: Prompt template for the SafeRx-Agent VERIFY operator described in Section 4.4. The system prompt encodes the retain/remove adjudication policy; the user prompt is dynamically assembled from matrix-retrieved DDI and contraindication flags, drug-name lookups, and prior-medication status. A `[PRIOR]` tag marks medications carried over from a prior visit so the model can bias toward continuation. The output schema is closed under `{RET, REM}`, matching Algorithm 1.

Direct LLM Prompt

System prompt:

You are an inpatient physician. For an ICU inpatient admission, predict the ATC Level-4 medication classes that will appear on this patient's medication list during the admission.

User prompt:

`{patient_text}`

Task instruction:

Think step by step, then emit a single JSON object. ATC L4 codes are exactly 5 characters: letter + 2 digits + 2 letters (e.g., `C10AA`, `B01AB`, `N02BE`). Do not invent codes, if unsure, omit.

Output format:

Output STRICT JSON only (no markdown, no commentary):

```
{
  "reasoning": "<step-by-step reasoning, plain text>",
  "predicted_drugs": [
    {"code": "<ATC L4>"}
  ]
}
```

Figure 11: Prompt template for the direct prompting baseline.

General-Agent Prompt

Note. The general-agent baseline reuses SafeRx-Agent’s SUMMARIZE and GENERATE prompt structure (Figures 7 and 8) and the shared CRITIQUE prompt (Figure 9), but replaces the routed specialty experts with a single *general_agent* skill that covers every therapeutic domain. Only the skill’s injected *summarization playbook* and *drug-prediction checklist* differ from the SafeRx-Agent versions; both are reproduced below.

SUMMARIZE stage — general-agent playbook:

Summarize the patient record to support broad inpatient medication prediction. The summary should capture active conditions, relevant chronic history, procedures, acuity of care, medication-relevant risks, and prior medication evidence. It should be comprehensive across therapeutic domains rather than specialized to one organ system.

In particular, the summary should preserve information that may affect medication choice, including acute diagnoses, comorbidities, procedures, critical-care status, infection evidence, pain or neurologic needs, renal and metabolic abnormalities, gastrointestinal or nutritional status, VTE or bleeding risk, and medications that may require continuation from prior visits.

GENERATE stage — general-agent drug-prediction instruction:

Predict ATC-L4 medication classes for the current inpatient admission as a single generalist agent. Use the patient summary, prior-visit medication evidence, diagnosis-linked medication evidence, and the closed ATC-L4 vocabulary. The prediction should cover both disease-directed therapy and common inpatient supportive care, while requiring patient-specific evidence for each medication class.

The agent should reason over the case by identifying active problems and chronic conditions, mapping each to plausible medication needs, checking whether prior medications should be continued, and considering medications associated with procedures, critical illness, infection management, pain control, thrombosis prevention, gastrointestinal protection, electrolyte correction, nutrition, and bowel care. Medication classes should be included only when supported by the current visit context, longitudinal history, or explicit evidence in the record.

The agent should prefer adequate coverage of clinically supported medication needs, but should avoid adding classes solely because they are common in inpatient care. Confidence should reflect the strength of evidence: higher confidence is reserved for explicitly documented medications or indications, while lower confidence is used for plausible but indirect clinical support.

Figure 12: Prompt template for the general-agent baseline. The baseline reuses SafeRx-Agent’s SUMMARIZE, GENERATE, and CRITIQUE structures, but replaces routed specialty experts with one all-domain *general_agent*.

RareAgents-Style Prompts Adapted to ATC-L4

Stage 1: Attending Physician. You are the Attending Physician coordinating a multi-disciplinary team (MDT). Given the patient profile and a pool of 41 clinical specialists, each defined by an id, name, and clinical scope, select the specialists whose expertise is relevant to the patient’s active diagnoses, procedures, and medication needs. Prefer disease- and organ-system specialists when they directly match the case, and include cross-cutting specialists such as clinical pharmacy or internal medicine when their scope can improve treatment review. Return the selected specialist IDs and a brief rationale as JSON.

Stage 2: Specialist Discussion. You are the {specialist_name} specialist in the MDT. Based on the patient profile, prior prescription history, and your specialty scope, discuss which medication classes are clinically relevant for the current admission. Use the provided DrugBank and DDI-graph feedback, when available, to note treatment evidence and potential safety concerns. Propose only medications within your specialty scope. Output 5-character ATC-L4 codes adapted to our closed prediction vocabulary, and convert ingredient-level ATC codes to their L4 parent. Return an empty list if your specialty has no relevant role. Return the proposed codes and reasoning as JSON.

Stage 3: Attending Synthesis. You are the Attending Physician finalizing the medication recommendation after the MDT discussion. Integrate the specialist proposals, prior prescription evidence, and DrugBank/DDI-graph feedback to form the final prescription set. Use the specialist proposals as the candidate pool, keep medications supported by the patient context or prior-visit precedent, and remove unsupported or tool-flagged candidates when the evidence does not justify keeping them. Do not introduce medications that were not proposed by any specialist. Output 1–30 ATC-L4 codes with reasoning as JSON.

Figure 13: **RareAgents-style baseline prompts adapted to ATC-L4 medication recommendation.** We adapt the RareAgents MDT workflow (Chen et al., 2026) to our closed-vocabulary ATC-L4 setting. The three stages correspond to attending-led specialist selection, specialist discussion with DrugBank and DDI-graph feedback, and attending-led synthesis. Unlike the original RareAgents medication prompt, which selects medication names from a fixed candidate list, our adaptation requires 5-character ATC-L4 codes from the prediction vocabulary.

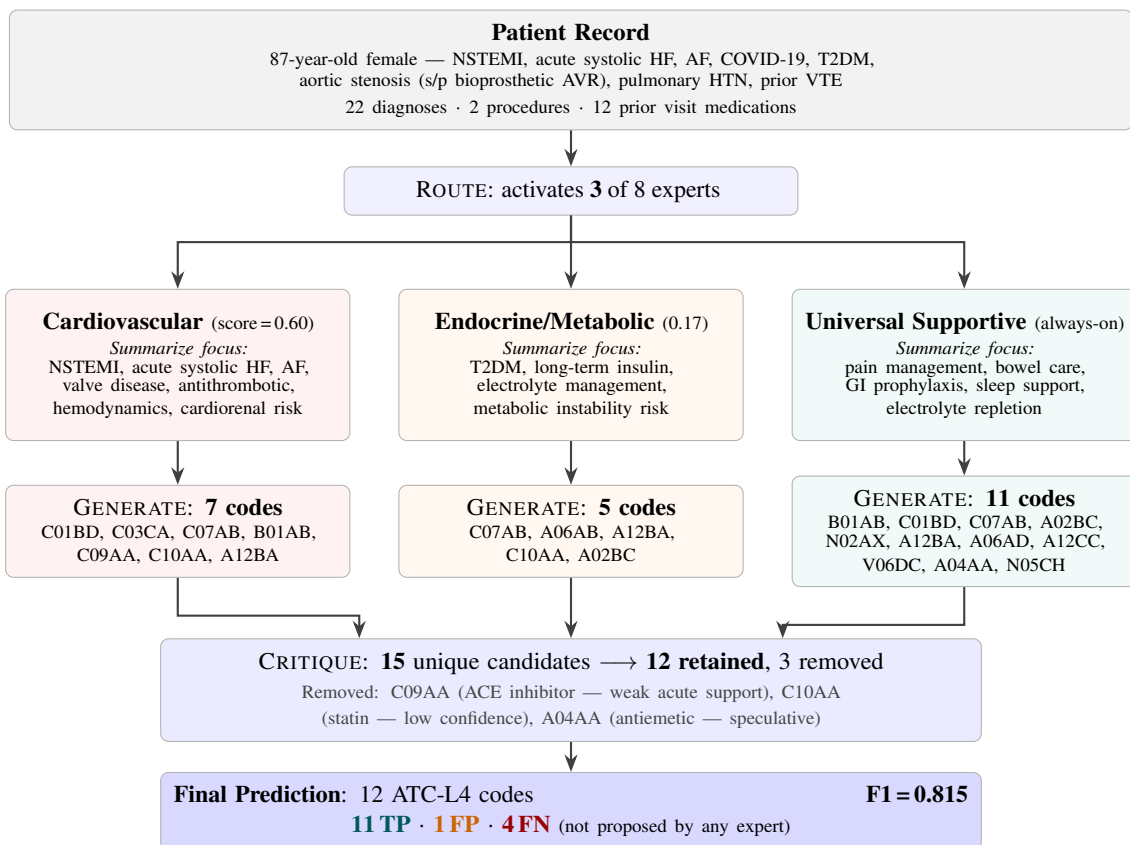


Figure 14: **Case study pipeline flow.** The patient record is routed to three specialty experts. Each expert summarizes the record from its domain perspective and generates ATC-L4 candidates. CRITIQUE merges the 15 unique candidates across all experts and removes 3 weakly supported codes, producing a final set of 12 medications.