PRISMATIC : Prescription Risk Inspection System for Multi-Agent Tactical Interaction in Clinical Decision

Anonymous ACL submission

Abstract

Medication prescribing errors remain a critical challenge in clinical practice, often stemming from incomplete patient understanding, ambiguous documentation, and suboptimal decision support. In this paper, we propose **PRISMATIC** [1], a 3-layer multi-agent prescription risk mitigation framework designed to generate safe, interpretable, and traceable drug regimens by analyzing unstructured patient clinical note texts. To enhance adaptability and safety, **PRISMATIC** integrates two mechanisms: (1) **Dynamic Self-updating Weight Adjustment (DSWA)**, which tunes risk factor weights over time, and (2) **Differential Feedback Calibration Mechanism(DFCM)**, which learns from discrepancies with gold-standard prescriptions to improve future outputs. Evaluated on a curated dataset from MIMIC-IV, PRISMATIC outperforms raw LLM outputs and prompting-based baselines (Few-Shot, Chain-of-Thought, ReAct, Tree-of-Thoughts) in reducing prescription risks. These results highlight the potential of multi-agent systems for improving clinical medication decision support.

1 Introduction

Medication prescribing plays a pivotal role in patient care but remains complex and error-prone. Prescribing decisions frequently arise from a synthesis of clinical guidelines and individual clinician judgment, resulting in significant variability, especially in challenging clinical contexts. As shown by the 33 influencing factors identified by (Davari et al., 2018), this variability can result in suboptimal or harmful prescriptions. The problem is widespread: (Alqenae et al., 2020) reported that nearly 1 in 5 adults experience adverse drug events post-discharge, while (Camacho et al., 2024) estimated 10,000 errors per 100,000 admissions in England, highlighting the urgent need for smarter prescribing support tools.

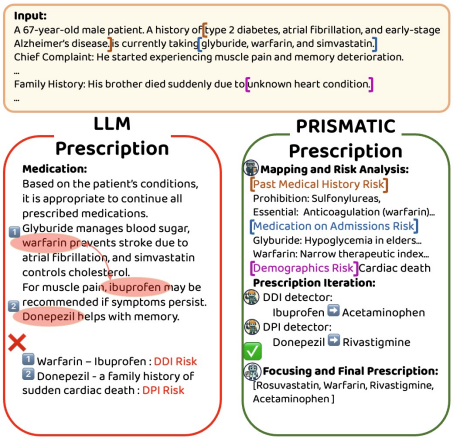


Figure 1: LLM Prescription vs. **PRISMATIC** Prescription

Early prescribing support tools relied on static rules (e.g., drug-drug interactions, contraindications) or manual chart reviews (Segura-Bedmar et al., 2010, 2011). Recent NLP and deep learning models automate tasks like ADE detection and medication extraction (Siegersma et al., 2022; Mashima et al., 2022), but remain fragmented, focusing on isolated tasks and lacking a holistic understanding of nuanced, unstructured clinical narratives such as symptoms, allergies, or evolving histories.

On the other hand, Large Language Models (LLMs) have recently shown human-level capabilities in reasoning and planning, spurring interest in healthcare applications (Thirunavukarasu et al., 2023). Studies have applied LLMs to streamline clinical workflows (Low et al., 2025), assist with prescribing and diagnosis (Kim et al., 2024; Chen et al., 2025a; Pan et al., 2025), and improve patient comprehension (Hsu et al., 2025; Hao et al., 2024). However, the use of LLM for drug prescribing support, a domain characterized by complex reasoning over dynamic patient-specific data, remains relatively underexplored. As agent-based system design evolves, it becomes increasingly feasible to

[1] Our implementation of PRISMATIC is available at https://anonymous.4open.science/r/PRISMATIC.

envision collaborative, LLM-powered multi-agent workflows that more effectively integrate diverse patient data and clinical knowledge to improve prescribing accuracy, safety, and personalization.

Inspired by the above, in this paper, we introduced **PRISMATIC**, a collaborative multi-agent architecture leveraging patient statements, drug instructions, and clinical knowledge for prescription risk inspection.

We evaluate **PRISMATIC** on the combined **MIMIC-IV Note** (Johnson et al., 2023) and **MIMIC-IV Hosp** (Johnson et al., 2024) datasets against raw LLM outputs and strong prompting baselines (Few-Shot, Chain-of-Thought (CoT)(Wei et al., 2023), ReAct(Yao et al., 2023b), Tree-of-Thoughts (ToT)(Yao et al., 2023a)). Empirical results (Figure 1) show that **PRISMATIC** consistently outperforms all baselines in resolving prescribing conflicts while enhancing safety, interpretability, and traceability.

To summarize, our main contributions are as follows:

- We introduce a multi-agent system, **PRISMATIC**, that leverages patient clinical text and clinical knowledge to perform prescription risk checks, assist in drug decision-making, and generate safer, lower-risk prescriptions.
- We introduce two mechanisms: **Dynamic Self-updating Weight Adjustment (DSWA)** and **Difference Feedback Calibration Mechanism (DFCM)** for self-adaptive risk modeling and iterative refinement.
- Through experiments, we demonstrate the decent performance of the **PRISMATIC** system in detecting and resolving prescription conflicts compared to both raw LLM outputs and state-of-the-art prompting engineering baselines.

2 Related Works

2.1 Multi-Agent System in Medications

Multi-agent systems have long been explored in healthcare due to their decentralized, modular nature, which enables distribution of specialized tasks and supports dynamic decision-making in complex clinical settings.

As one of the early explorations nearly two decades ago, (Rodríguez et al., 2005) proposed a rule-based agent framework to support doctor–patient collaboration and personalized hospital assistance. A decade later, (Benhajji et al., 2015) introduced a multi-agent system for managing patient flow and hospital resource allocation.

More recently, with the rapid advancement of AI and large language models (LLMs),multi-agent systems have evolved significantly, overcoming prior limitations in perception and interaction (Li et al., 2024). Recent systems leverage LLMs to enhance clinical decision-making (Chen et al., 2025b), support surgical workflows with chain-of-thought reasoning (Low et al., 2025), and enable collaborative diagnostic reasoning among doctor agents (Chen et al., 2025a). Others incorporate verified knowledge tools (Gao et al., 2025) or adaptive frameworks mimicking real-world clinical decision-making (Kim et al., 2024). These developments underscore the growing sophistication and promise of LLM-powered multi-agent systems in improving healthcare delivery.

These advancements highlight the increasing sophistication of LLM-based multi-agent systems in healthcare, demonstrating their potential to enhance decision-making processes and improve clinical outcomes across various medical domains.

2.2 Retrieval-Augmented Generation (RAG) in Medication Recommendation

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by retrieving relevant knowledge from external sources to inform generation, improving factual accuracy, explainability, and reducing hallucinations (Lewis et al., 2021; Gao et al., 2024; Shuster et al., 2021).

Considering the medical domain, where accuracy and reliability are paramount, RAG has shown promise in clinical question answering, guideline-based support, and evidence-grounded summarization (Sohn et al., 2024; Lu et al., 2024; Lopez et al., 2025). Several studies have used structured sources, such as drug labels, clinical guidelines, and biomedical literature, to enhance generation. For example, MedRAG (Zhao et al., 2025) integrates LLMs with DrugBank, UMLS, and PubMed to improve the safety and factual precision of medical recommendations.

3 Preliminary

3.1 Problem Definition

We consider the task of generating a safe and interpretable prescription from unstructured clinical text.

Input: Clinical Statement where each t_i denotes a

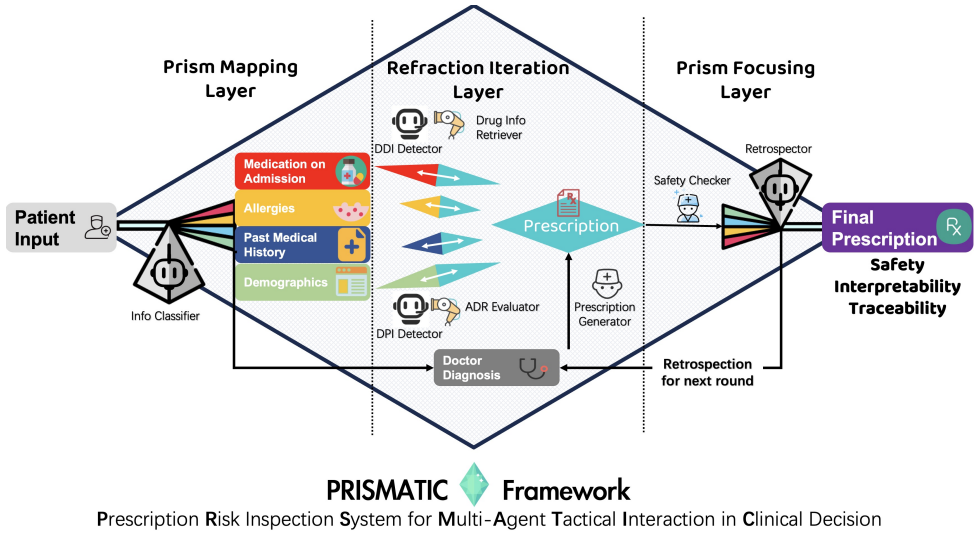


Figure 2: **PRISMATIC** Multi-Agent System Framework

segment of the patient's unstructured clinical notes (e.g., medications on admission, family history).

$$T = \{t_1, t_2, \ldots, t_n\}$$

Output: Prescription

$$P = \{(d_i, u_i, r_i, e_i)\}_{i=1}^{N}$$

Here, $d_i \in \mathcal{D}$ denotes a selected drug from the formulary, u_i is its dosage plan, r_i is the administration route, and e_i is a human-readable explanation. To solve this problem, the following core factors must be introduced to control risks and enhance the rationality and safety of drug use:

1. Interactions between drugs: $DDI(d_i, d_j)$
2. Interactions between drugs and patient information: $DPI(d_i, T)$
3. Validation of dosage, route and explanation: $Check_{u_i}$/$Check_{r_i}$/$Check_{e_i}$

To mitigate prescription risks and prevent medication errors, our system is designed to generate prescriptions that are safe, interpretable, and traceable. To this end, we propose that each generated prescription must satisfy the following criteria:

- *Safety:*
  $\forall\, i \neq j:\ DDI(d_i, d_j) = 0$
  $\forall\, i:\ DPI(d_i, T) = 0.$
- *Interpretability:*
  $\forall\, i:\ Check_{e_i} = 0$
  Each explanation e_i must compliant with relevant clinical guidelines and clearly articulate the rationale for selecting d_i.
- *Traceability:*
  $\forall\, i:\ Check_{u_i} = 0, \forall\, i:\ Check_{r_i} = 0$
  Dosage u_i and administration route r_i must be verifiable, and the entire decision-making process must be logged for audit.

3.2 LLM-based Prescription Generation

As a baseline, we implement a direct LLM-based approach, where the entire process is treated as an end-to-end mapping without any intermediate analysis or structured reasoning. Formally, this can be represented as:

$$\mathcal{F} : T \longmapsto \mathcal{P}$$

The input clinical text T is provided to a general-purpose language model in the form of a prompt, and the final prescription $\mathcal{P}$ is generated directly. Various prompt engineering techniques and reasoning strategies (e.g., CoT, ToT) are applied to optimize the output.

4 Proposed Approach – PRISMATIC

To reduce prescription risks and prevent medication errors in clinical adjuvant drug decision-making, we propose **PRISMATIC**, a three-layer multi-agent tactical interaction system, as illustrated in Figure 2. The system is inspired by the behavior of a prism: just as a prism decomposes white light into distinct spectral components and then recombines them into a coherent beam, **PRISMATIC** decomposes clinical input into specialized dimensions, refines each through agent interactions, and integrates the results to produce a safe and informed prescription.

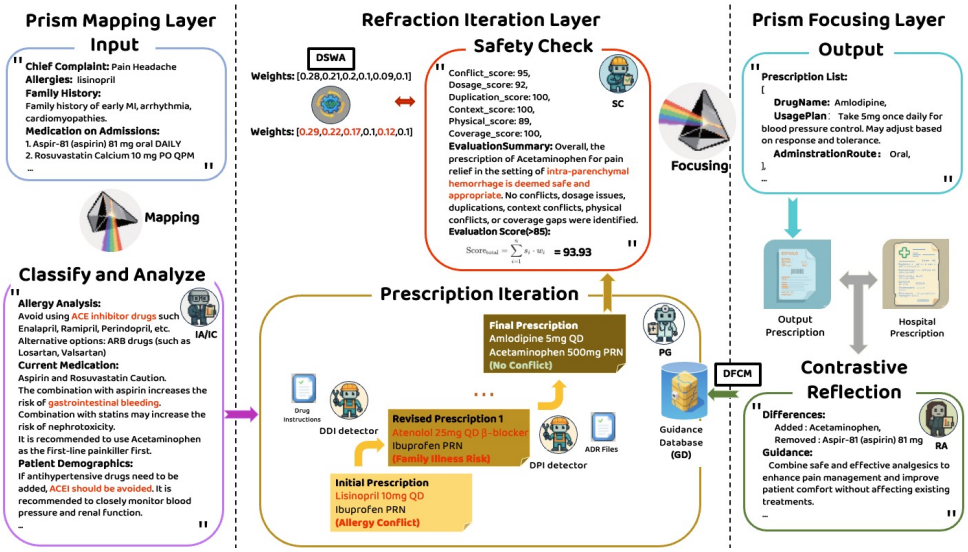


Figure 3: **PRISMATIC** Multi-Agent System Workflow

The **prism mapping layer** decomposes unstructured clinical notes into multiple safety-critical aspects, including demographics, allergic history, medication history, and medications on admissions. Then each aspect is analyzed by specialized agents through parallel reasoning. In the **refraction iteration layer**, these agents continuously interact with the prescribing agent, refining their recommendations in an iterative process that mirrors light refraction, gradually reducing prescription risk and improving decision quality. Once the aggregated safety score (analogous to a refractive index) exceeds a predefined threshold, the system proceeds to the **prism focusing layer**, where the refined outputs are synthesized into a final, interpretable, and traceable prescription, much like refracted light converging into a coherent beam. Formally, let the input be

$$T = \{t_1, t_2, \ldots, t_n\}$$

where each t_i denotes different aspects of the patient's unstructured clinical notes.
Our goal is to learn through a mapping layer, an iteration layer, and a focusing layer:

$$\mathcal{F}_{\text{mapping}} : T \longmapsto \mathcal{A} = \{a_1, a_2, \ldots, a_n\}, \mathcal{P}_{\text{ini}}$$

$$\mathcal{F}_{\text{iteration}} : \mathcal{A}, \mathcal{P}_{\text{ini}} \longmapsto \mathcal{P}_{\text{n}}$$

$$\mathcal{F}_{\text{focusing}} : \mathcal{P}_{\text{n}} \longmapsto \mathcal{P}_{\text{final}}$$

where the multiple facets that affect the safety of the prescription is defined as:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$$

the output prescription $\mathcal{P}_{\text{final}}$ is defined as:

$$\mathcal{P}_{\text{final}} = \{(d_i, u_i, r_i, e_i)\}_{i=1}^{k}$$

Here, $d_i \in \mathcal{D}$ denotes a selected drug from the formulary, u_i is its dosage plan, r_i is the administration route, and e_i is a human-readable explanation.

4.1 PRISMATIC Framework

4.1.1 Prism Mapping Layer

Input Structuring and Analyzing. The prism mapping layer is used to extend the mapping of the patient's input information to each structured factor edge and analyze the potential medication risks that each factor may cause. There are two agents in this layer:

- **Information Cleaner Agent(IC).** IC cleans the patient information and classifies it into various dimensions. In our architecture, we classify the text information into four dimensions:

$$\mathcal{F}_{\text{IC}} : \mathcal{T} \longmapsto \mathcal{T}' = \{t_{BDI}, t_{AH}, t_{PMH}, t_{MOA}\}$$

  - Basic Demographics Information (BDI)
  - Allergic History (AH)
  - Past Medical History (PMH)
  - Medications on Admission (MOA)
- **Information Analyst Agent(IA).** IA analyzes the categorized information $\mathcal{T}'$ and output the different aspects $\mathcal{A}$ of potential risks and dangerous conflicts that each type of information may trigger for reference in the subsequent

agent analysis.

$$\mathcal{F}_{\text{IA}} : \mathcal{T}' = \{t_{BDI}, t_{AH}, t_{PMH}, t_{MOA}\} \longmapsto \mathcal{A} = \{a_{BDI}, a_{AH}, a_{PMH}, a_{MOA}\}$$

4.1.2 Refraction Iteration Layer

Prescription Generation and Conflict Inspection. Using the structured profile $\mathcal{T}'$ and risk $\mathcal{R}$ as input, this layer iteratively constructs, evaluates, and refines candidate prescriptions by simulating multi-agent interactions. The goal is to resolve all known drug-drug and drug-patient conflicts through iterative feedback. Agents tactically collaborate through repeated "refraction" cycles, until a stable, safe solution is reached. Key agents include:

- **Prescription Generator (PG).** Given the multi-dimensional patient profiles from the mapping layer, PG prescribes through the guidance of clinical guidelines and rule databases. The **Guidance Database (GD)** is updated from the content generated by each round of backtracking and reflection.

$$\mathcal{F}_{\text{PG}} : \mathcal{T}'/\mathcal{R} \xrightarrow{\text{GD}} \mathcal{P}$$

- **DDI/DPI Detector (DDI/DPI).** DDI/DPI detect potential risks in drug–drug interactions (DDI) and drug–patient interactions (DPI). We use **Retrieval-Augmented Generation(RAG)** to leverage the instructions of the drugs in DrugBank Knowledge files. It returns a detailed conflict report, including risk levels and explanations.

$$\mathcal{F}_{\text{DDI/DPI}} : \mathcal{P}_n \xrightarrow{\text{RAG}} \mathcal{R}_{\text{conflict}} = \{(d_i, d_j, s_{ij}, e_{ij})\} \longmapsto \mathcal{P}_{n+1}$$

  - $\mathcal{R}_{\text{conflict}}$ =
    - d_i, d_j: Interaction drugs
    - s_{ij}: Interaction level
    - e_{ij}: Explanation for interaction

4.1.3 Prism Focusing Layer

Once the Refraction Iteration Layer produces a regimen whose safety score meets or exceeds the convergence criterion, the **Prism Focusing Layer** performs final validation and convergence of the prescription, ensuring all checks passed and explanations attached. It employs two specialized agents:

- **Safety Checker (SC).** SC conducts the final evaluation of the prescription P, scoring it across drug conflict score, dosage score, drug duplication score, patient information score, administration routes score, and drug coverage score, six risk dimensions using the **Dynamic Self-updating Weight Adjustment (DSWA)** mechanism (see Section 4.2):
  The SC function is defined as:

$$\mathcal{F}_{\text{SC}} : \mathcal{P} \longmapsto Score$$

$$w_i' = \frac{e^{s_i}}{\sum_{j=1}^{6} e^{s_j}}, \quad \text{for } i = 1, \ldots, 6$$

  The detailed dynamic self-updating algorithm is displayed in Algorithm 1.

- **Retrospection Agent (RA).** RA reviews the generation process using the **Differential Feedback Calibration Mechanism (DFCM)** (see Section 4.3), comparing the final output P_{final} with the ground-truth P_{gt}, analyzing differences, and updating the guidance database to refine future outputs from the Prescription Generator (PG).

By "focusing" the multi-faceted outputs of the preceding layers, the Prism Focusing Layer produces a single, optimized prescription that is safe, interpretable, and fully traceable from initial input to final recommendation.

4.2 Dynamic Self-updating Weight Adjustment (DSWA)

To enable adaptive learning and stable convergence in multi-agent collaboration, we propose a **Dynamic Self-updating Weight Adjustment (DSWA)** mechanism. DSWA allows agents to iteratively adjust their influence based on prescription risk signals and performance feedback.
Prescription risks are grouped into six dimensions: drug conflict, dosage, duplication, clinical context, administration route, and insurance coverage. Each is initially weighted based on empirical frequency and clinical severity from (Friedman et al., 2007):

$$\boldsymbol{\omega}^{(0)} = \begin{bmatrix} \omega_{\text{conflict}}, & \omega_{\text{dosage}}, & \omega_{\text{duplication}}, \\ \omega_{\text{context}}, & \omega_{\text{administration}}, & \omega_{\text{coverage}} \end{bmatrix} = \begin{bmatrix} 0.35, & 0.26, & 0.15, \\ 0.10, & 0.12, & 0.02 \end{bmatrix}$$

This initial weight vector guides the Safety Checker (SC) in evaluating risk dimensions. Based on feedback from intermediate prescriptions and identified

risk patterns, DSWA then updates these weights iteratively. The adjustment process takes into account the marginal contribution of each dimension to overall risk, enabling the system to self-correct and better prioritize critical issues. The following is the detailed algorithm:

Algorithm 1 Dynamic Self-updating Weight Adjustment

Require: Current weights $\omega^{(t)}$, scores $\mathbf{s}$, smoothing α, temperature β
Ensure: Updated weights $\omega^{(t+1)}$

Step 1: Compute raw weights via softmax
for $i = 1$ to 6 **do**
  $\tilde{\omega}_i \leftarrow \exp\big(\beta\, s_i\big)$
end for
$Z \leftarrow \sum_{j=1}^{6} \tilde{\omega}_j$
for $i = 1$ to 6 **do**
  $\omega_i^{\text{new}} \leftarrow \tilde{\omega}_i / Z$
end for
Step 2: Exponential smoothing fusion
for $i = 1$ to 6 **do**
  $\omega_i^{(t+1)} \leftarrow \alpha\, \omega_i^{(t)} + (1-\alpha)\, \omega_i^{\text{new}}$
end for
Return $\omega^{(t+1)}$ =0

4.3 Differential Feedback Calibration Mechanism (DFCM)

To better align with clinical standards and improve prescription quality, we propose the **Differential Feedback Calibration Mechanism (DFCM)**. At each iteration, the system compares its output $\mathcal{P}$sys with the gold-standard hospital prescription $\mathcal{P}$gt. DFCM identifies discrepancies in drug choice, dosage, and administration, traces their root causes, and encodes corrective heuristics into a centralized **Guidance Database (GD)**. These rules refine the Prescription Generator in future rounds, reducing repeated errors and guiding convergence toward clinically approved patterns.

5 Experiments

5.1 Experiment Setup

Evaluation Datasets. We evaluate **PRISMATIC** using a custom *clinical note–prescription* dataset built from MIMIC-IV:

- **Data Filtering and Linking.** We link the `mimiciv_note` and `mimiciv_hosp` tables via the unique patient identifier `subject_id`, ensuring that each clinical note is matched with the corresponding hospital record. From the `diagnoses_icd` table (ICD-10 version), we select hospital admissions with 3–8 chronic conditions (from `diagnoses_icd`) and 5–20 medications (from `prescriptions`) to ensure moderate case complexity. admissions with 3-8 chronic conditions to ensure moderate complexity of the patient's condition.
- **Note–Prescription Pairing.** Drug names are normalized using RxNav with RxNorm terms (U.S. National Library of Medicine, 2025) into RxCUI. Discharge summaries are then paired with prescriptions via `subject_id` and `hadm_id`. That forms the **CCM Dataset** (Compound Condition Medication Dataset).
- **Dataset final results.**
  - *subject_id*: Patient's unique identifier.
  - *text*: Unstructured patient clinical text.
  - *prescriptions:* A ground-truth list of drugs, including drug RXCUI code, dosage and administration route.

The final CCM Dataset includes 5,375 matched note–prescription pairs for evaluation. **Evaluation Metrics** We consider evaluating the performance with the following metrics.

- **Overlap Rate (OR)**. The overlap rate measures the degree of coverage between the output prescription drug array P_i and the ground truth prescription label $P_{gt}i$. For the i-th case:

$$\text{OR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\text{P}_i \cap \text{P}_{gt}i\right|}{\left|\text{P}_{gt}i\right|}$$

- **Precision (Prec.)**. The accuracy rate measures what proportion of the prescription drug array P_i actually need to be prescribed. For the i-th case:

$$\text{Prec.} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\text{P}_i \cap \text{P}_{gt}i\right|}{\left|\text{P}_i\right|}$$

- **Exact Match Ratio (EM)**. The degree of perfect match refers to the percentage of completely correct prescriptions in the total cases.

$$\text{EM} = \frac{P_{correct}}{n}$$

Baseline Methods.

- Zero-Shot: A single general-purpose LLM without prompt engineering, framework, or external knowledge.

- Few-Shot: Uses a few in-context examples to guide the model in task understanding and execution.
- ReAct: Integrates reasoning and action, allowing the model to think first and then action.
- Chain-of-Thought: Promotes step-by-step reasoning before reaching a final answer.
- Tree-of-Thought: Builds on Chain-of-Thought by enabling exploration of multiple reasoning paths in a tree structure.

Model	Method	OR	Prec.	EM
GPT-4o	Zero-Shot	29.11	25.72	2.05
	CoT	37.40	44.30	5.51
	ToT	42.81	50.54	7.22
	ReAcT	38.63	41.11	5.12
	Few-Shot	31.61	29.09	2.57
	PRISMATIC (Ours)	**56.81**	**60.11**	**13.58**
Llama-3.1-8B	Zero-Shot	24.50	28.10	2.23
	CoT	32.40	30.22	4.89
	ToT	45.86	49.03	7.66
	ReAcT	39.63	41.98	7.81
	Few-Shot	26.61	32.09	5.38
	PRISMATIC (Ours)	**51.40**	**56.70**	**10.44**

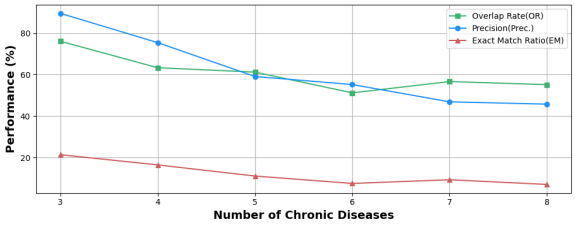
Table 1: Comparison of Different Methods on GPT-4o and Llama-3.1-8B-Instruct

5.2 Main Results

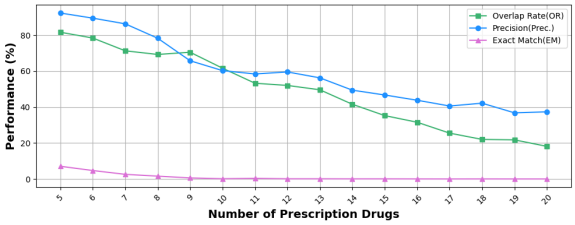
The main results of our experiments on the CCM dataset using two models (GPT-4o and Llama-3.1-8B-Instruct) are shown in Table 1. Several key findings emerge: **First**, the proposed **PRISMATIC** framework consistently achieves the best performance across all metrics and both models. In particular, with GPT-4o, it attains an overlap rate of 56.81%, a precision of 60.11%, and an exact match (EM) of 13.58%. Similar trends are observed with Llama-3.1-8B-Instruct, confirming the model-agnostic advantage of our multi-agent approach. **Second**, among baseline methods, Tree-of-Thought performs best. Its strategy of generating and evaluating multiple prescription plans yields higher medication diversity and quantity, leading to improved coverage and overlap metrics. **Third**, all methods exhibit low EM scores, with the best reaching only 14%, underscoring the persistent gap between LLM-generated prescriptions and human clinical standards. These results demonstrate that **PRISMATIC** significantly enhances prescription generation performance over standard prompting methods (Zero-Shot, CoT, ToT, ReAct, Few-Shot).

5.3 Quantitative Analysis

Task Complexity. To further understand the performance limitations, we assess prescription accuracy as patient complexity increases along two axes—number of chronic conditions and number of ground-truth drugs—shown in Figure 4 and Figure 5. Using **PRISMATIC** with GPT-4o as an example, we observe several consistent patterns.



(a) Performance vs. Number of Chronic Diseases



(b) Performance vs. Number of Prescription Drugs

Figure 4: Performance trends of different methods across varying levels of patient and prescription complexity. (a) Performance versus the number of chronic conditions per patient. (b) Performance versus the No. medications in the ground-truth prescription.

Precision > Overlap Rate. Across all complexity levels, precision consistently exceeds overlap Rate. This suggests that when a drug is recommended by the system, it is often correct. However, the model frequently fails to cover all necessary medications, indicating limited recall or incomplete coverage of the full prescription.

Performance Declines with Complexity. As either the number of chronic conditions or the num-

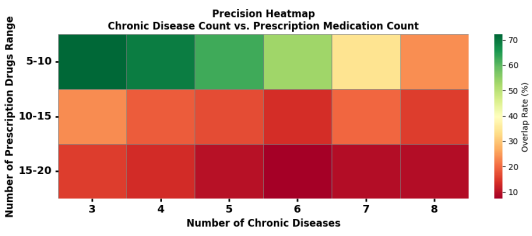


Figure 5: Heatmap of precision across varying chronic disease and prescription complexities.

ber of target drugs increases, model performance declines across all metrics. This reflects increased clinical complexity, where more comorbidities and therapeutic demands lead to more difficult prescription decisions.

Exact Match is Rare. When prescriptions include over 10 drugs, achieving a complete match becomes nearly impossible.

5.4 Ablation Study

To assess the contributions of key components in our framework, we conduct an ablation study on four modules: DSWA, DFCM, and the DDI/DPI detectors. **Removing DSWA and DFCM**—used in the Safety Checker and Retrospector—leads to noticeable drops in Precision and Overlap Rate. **Disabling DDI/DPI detectors** results in a more substantial performance decline and a sharp rise in potential risk cases. These detectors, powered by *DrugBank* via RAG, are critical for aligning prescriptions with safety standards. As shown in Table 2, **PRISMATIC** consistently outperforms the ablated variants, underscoring the importance of both interaction detection and iterative refinement.

Framework	OR	Prec.	EM
w/o DSWA	48.55	55.11	11.56
w/o DFCM	49.56	59.22	13.42
w/o DDI detector	41.25	50.22	8.56
w/o DPI detector	44.13	49.65	7.33
PRISMATIC	**56.81**	**60.11**	**13.58**

Table 2: Ablation study of PRISMATIC by removing each module individually.

5.5 Error Analysis

Our Safety Checker Agent (SC) generates a reflection document after each assessment, identifying safety issues in the final prescription based on four error categories: Basic Demographic Information (BDI), Allergic History (AH), Past Medical History (PMH), and Medications on Admission (MOA), as shown in Figure 6.

Among these, allergy-related risks were minimal,

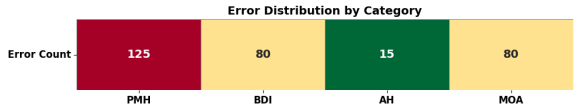


Figure 6: Error Distribution by Category

indicating effective handling of AH. In contrast, errors related to PMH and MOA were most frequent. **PMH errors** highlight the need for thorough review of conditions such as heart failure, liver, or kidney disease, which critically influence drug choice and risk of interactions. **BDI** also plays an important role in customizing treatment, with family history occasionally revealing hidden risks. **MOA errors**: including omissions, duplications, or inappropriate continuations, reflect challenges in accurate medication reconciliation, further underscoring the value of traceable and context-aware prescription generation. **??** shows how weights adjust over iterations across six risk dimensions. While initial rankings are mostly reasonable, dosage gradually becomes the most critical factor, overtaking drug–drug interactions. This suggests that dosage errors may play a larger role in real-world prescription safety.

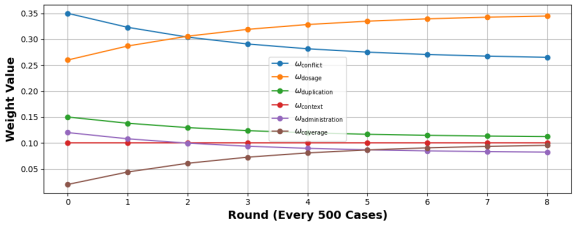


Figure 7: Evolution of risk weights across six dimensions in the DSWA mechanism. Every 500 cases per iteration round. The line graph illustrates how weights dynamically adjust over time.

6 Conclusion

In this work, we introduced **PRISMATIC**, a multi-agent collaboration system to generate safe, interpretable, and traceable drug regimens based on patient clinical note texts. By integrating dynamic feedback mechanisms: **DSWA** and **DFCM**, our system iteratively refines its knowledge base and prescription quality. Through layered agent collaboration, from data extraction to safety validation and final prescription, PRISMATIC creates a closed-loop learning process that bridges automated reasoning with clinical guidelines. Tested on MIMIC-IV dataset shows our agent system consistently outperforms raw LLMs and standard prompting methods, showcasing its effectiveness and applicability.

7 Limitations

This study has several limitations that define its scope and suggest future research directions:

- **Incomplete Clinical Scope:** he ground truth prescriptions used as references often include preemptive, supportive, or prognostic medications that address comorbidities, complication prevention, or long-term patient management. These prescriptions reflect complex clinical judgments extending beyond the primary diagnosis. However, our system primarily focuses on generating prescriptions directly related to the diagnosed condition, which may omit such broader therapeutic considerations routinely made by clinicians. This gap limits the system's ability to fully capture the holistic medication strategies used in real-world practice.
- **Limited Prescription Accuracy and Generalizability:** Although the system incorporates advanced mechanisms such as dynamic self-updating weights and differential feedback calibration to iteratively improve performance, the overall accuracy and alignment with expert prescriptions remain suboptimal, especially in complex cases involving multiple conditions and medications. The prescription generator currently struggles to precisely select optimal drugs, dosages, and administration routes in diverse clinical scenarios. Moreover, the system's performance is constrained by the scope and richness of the medical knowledge integrated. Enhancing domain coverage with more comprehensive clinical guidelines, drug databases, and real-world practice patterns is necessary to increase robustness and clinical applicability.
- **Evaluation Constraints:** Our evaluation relies heavily on retrospective datasets and reference prescriptions, which may not fully represent real-time clinical decision-making dynamics or patient-specific nuances. The absence of prospective validation in live clinical settings restricts our ability to assess the system's practical utility and safety in everyday healthcare environments.

8 Acknowledgment

The authors thank anonymous colleagues for their helpful discussions. This work received no external funding.
The authors used AI-assisted writing tools for language editing purposes only. All content and ideas were developed by the authors.

9 Ethics

This study does not involve direct experimentation on human or animal subjects. All data used were either publicly available or properly anonymized to ensure that no personally identifiable information (PII) was involved. This study uses the MIMIC-IV database, a publicly available, de-identified dataset of critical care patients. Access to MIMIC-IV is governed by the PhysioNet Credentialed Health Data License 1.5.0, which permits non-commercial research use under strict conditions to protect patient privacy. All authors complied with the data usage agreement and completed the required human subjects research training. No attempt was made to re-identify any individuals. The use of this dataset adheres to relevant ethical guidelines and institutional standards.

References

Fatema A. Alqenae, Douglas Steinke, and Richard N. Keers. 2020. Prevalence and nature of medication errors and medication-related harm following discharge from hospital to community settings: A systematic review. *Drug Safety*, 43(6):517–537.

N. Benhajji, D. Roy, and D. Anciaux. 2015. Patient-centered multi agent system for health care. *IFAC-PapersOnLine*, 48(3):710–714. 15th IFAC Symposium onInformation Control Problems inManufacturing.

Elizabeth M Camacho, Sean Gavan, Richard Neil Keers, Antony Chuter, and Rachel Ann Elliott. 2024. Estimating the impact on patient safety of enabling the digital transfer of patients' prescription information in the english nhs. *BMJ Quality & Safety*, 33(11):726–737. Published online 2024 Oct 18; PMC11503046.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025a. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8(1):159.

Zhen Chen, Zhihao Peng, Xusheng Liang, Cheng Wang, Peigan Liang, Linsheng Zeng, Minjie Ju, and Yixuan Yuan. 2025b. Map: Evaluation and multi-agent enhancement of large language models for inpatient pathways. *Preprint*, arXiv:2503.13205.

Majid Davari, Elahe Khorasani, and Bereket Molla Tigabu. 2018. Factors influencing prescribing decisions of physicians: A review. *Ethiopian Journal of Health Sciences*, 28(6):795–804.

Amy L. Friedman, Sarah R. Geoghegan, Noelle M. Sowers, Sanjay Kulkarni, and Richard N. Jr Formica.

2007. Medication errors in the outpatient setting: classification and root cause analysis. *Archives of Surgery*, 142(3):278–283.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Txagent: An ai agent for therapeutic reasoning across a universe of tools. *Preprint*, arXiv:2503.10970.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Yuexing Hao, Jason Holmes, Mark Waddle, Nathan Yu, Kirstin Vickers, Heather Preston, Drew Margolin, Corinna E. Löckenhoff, Aditya Vashistha, Marzyeh Ghassemi, Saleh Kalantari, and Wei Liu. 2024. Outlining the borders for llm applications in patient education: Developing an expert-in-the-loop llm-powered chatbot for prostate cancer patient education. *Preprint*, arXiv:2409.19100.

Hsin-Ling Hsu, Cong-Tinh Dao, Luning Wang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Chun-Chieh Liao, Pengfei Hu, Xiaoxue Han, Chih-Ho Hsu, Dongsheng Luo, Wen-Chih Peng, Feng Liu, Fang-Ming Hung, and Chenwei Wu. 2025. Medplan:a two-stage rag-based system for personalized medical plan generation. *Preprint*, arXiv:2503.17900.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). Accessed: 2025-05-19.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). Accessed: 2025-05-19.

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. *Preprint*, arXiv:2404.15155.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P. Ma, April S. Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, Nigam H. Shah, and Jonathan H. Chen. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhitao Zeng, Zhu Zhuo, Evangelos B. Mazomenos, and Yueming Jin. 2025. Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. *Preprint*, arXiv:2503.10265.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68, Bangkok, Thailand. Association for Computational Linguistics.

Yukinori Mashima, Takashi Tamura, Jun Kunikata, Shinobu Tada, Akiko Yamada, Masatoshi Tanigawa, Akiko Hayakawa, Hirokazu Tanabe, and Hideto Yokoi. 2022. Using natural language processing techniques to detect adverse events from progress notes due to chemotherapy. *Cancer Informatics*, 21:11769351221085064. PMID: 35342285.

Jie Pan, Seungwon Lee, Cheligeer Cheligeer, Elliot A Martin, Kiarash Riazi, Hude Quan, and Na Li. 2025. Integrating large language models with human expertise for disease detection in electronic health records. *Computers in Biology and Medicine*, 191:110161.

Marcela D. Rodríguez, Jesus Favela, Alfredo Preciado, and Aurora Vizcaíno. 2005. Agent-based ambient intelligence for healthcare. *AI Commun.*, 18(3):201–216.

Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. 2010. Extracting drug-drug interactions from biomedical texts. *BMC Bioinformatics*, 11(5):P9.

Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. 2011. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics*, 12(2):S1.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *Preprint*, arXiv:2104.07567.

Kim Siegersma, Martijn Evers, Sanne Bots, Floor Groepenhoff, Yolande Appelman, Leo Hofstra, Igor Tulevski, Gert-Jan L. Somsen, Hester den Ruijter, Maarten van Smeden, and Niek Onland-Moret. 2022. Development of a pipeline for adverse drug reaction identification in clinical notes: Word embedding models and string matching. *JMIR Medical Informatics*, 10(1):e31063.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *Preprint*, arXiv:2411.00300.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

U.S. National Library of Medicine. 2025. RxNorm API. https://lhncbc.nlm.nih.gov/RxNav/APIs/RxNormAPIs.html. Accessed: 2025-05-20.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *Preprint*, arXiv:2502.04413.

Appendix