

RES-MR: Risk-Aware Reasoning for Explainable and Safe Medication Recommendation

Cong Wang*

Key Laboratory of Computing Power
Network and Information Security,
Ministry of Education, Shandong
Computer Science Center (National
Supercomputer Center in Jinan), Qilu
University of Technology (Shandong
Academy of Sciences); Shandong
Provincial Key Laboratory of
Computing Power Internet and
Service Computing, Shandong
Fundamental Research Center for
Computer Science
Jinan, Shandong, China
cong.wang2024@gmail.com

Jin Li*

University of Technology Sydney
Sydney, NSW, Australia
jin.li-4@student.uts.edu.au

Shoujin Wang

University of Technology Sydney
Sydney, NSW, Australia
shoujin.wang@uts.edu.au

Yishuo Li

Key Laboratory of Computing Power
Network and Information Security,
Ministry of Education, Shandong
Computer Science Center (National
Supercomputer Center in Jinan), Qilu
University of Technology (Shandong
Academy of Sciences); Shandong
Provincial Key Laboratory of
Computing Power Internet and
Service Computing, Shandong
Fundamental Research Center for
Computer Science
Jinan, Shandong, China
yishuo.li@foxmail.com

Huilin Gu

University of Technology Sydney
Sydney, NSW, Australia
huilin.gu@student.uts.edu.au

Wenpeng Lu†

Key Laboratory of Computing Power
Network and Information Security,
Ministry of Education, Shandong
Computer Science Center (National
Supercomputer Center in Jinan), Qilu
University of Technology (Shandong
Academy of Sciences); Shandong
Provincial Key Laboratory of
Computing Power Internet and
Service Computing, Shandong
Fundamental Research Center for
Computer Science
Jinan, Shandong, China
wenpeng.lu@qlu.edu.cn

Abstract

Medication recommendations seek to deliver personalized drug combinations based on patients' clinical information, where ensuring transparent decision-making and personalized safety are of paramount importance. However, existing approaches predominantly operate as "black boxes", offering limited interpretability and hindering clinician trust. Furthermore, they typically rely on fixed,

patient-agnostic safety constraints, neglecting the heterogeneity of risk tolerance across individuals, which poses significant risks to vulnerable populations. To address these gaps, we propose **RES-MR**, a novel Risk-aware Reasoning framework for Explainable and Safe Medication Recommendation using LLMs. RES-MR follows a two-stage training paradigm: (1) an explainable clinical reasoning distillation stage to elicit diagnostic capabilities by fine-tuning on synthesized reasoning trajectories, grounded in patient-specific knowledge from graph retrieval; and (2) a risk-aware policy optimization stage to dynamically balance therapeutic efficacy with personalized safety. This stage employs disentangled drug factors to construct patient-specific reference prototypes that capture desired therapeutic drug features and risk features to avoid. These prototypes enable risk boundary calibration and guide the model toward safe recommendations via reward shaping and clipping. Extensive experiments on benchmark datasets demonstrate that RES-MR

*Equally contributed.

†Corresponding author.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or internal use, or the internal or personal use of specific clients, is granted by ACM for users registered with ACM, provided that the fee of \$12.00 is paid directly to ACM. This permission is granted without fee where organizations have been granted a special rate code by ACM. For all other use, permission should be sought from ACM. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '26, Melbourne, VIC, Australia.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2599-9/2026/07

<https://doi.org/10.1145/3805712.3809604>

significantly outperforms state-of-the-art baselines in terms of accuracy, safety, and explainability. Our code and data are available at: <https://github.com/wangcong2001/RES-MR>.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Social and professional topics** → **Medical records**.

Keywords

Medication recommendation, Electronic health record, Large language models, Explainability, Medication safety

ACM Reference Format:

Cong Wang, Jin Li, Shoujin Wang, Yishuo Li, Huilin Gu, and Wenpeng Lu. 2026. RES-MR: Risk-Aware Reasoning for Explainable and Safe Medication Recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3805712.3809604>

1 Introduction

Medication recommendation is a pivotal task in clinical decision support systems, aiming to assist physicians in tailoring effective and safe drug combinations for individual patients [1, 25, 53, 59, 60, 62, 64, 81]. Reliable medication recommender systems hold immense potential to not only mitigate preventable medication errors and reduce adverse patient outcomes but also significantly alleviate the cognitive load on clinicians [33, 36, 82]. However, in real-world clinical practice, this task is substantially more complex than following general recommendations, as accurate and safe medication decisions require jointly considering therapeutic benefit, potential drug-drug interactions, contraindications, and patient-specific context [14]. To address this, there is a growing paradigm shift towards utilizing data-driven approaches based on large-scale Electronic Health Records (EHRs) [43, 69, 72] that capture the rich longitudinal patient data necessary to model these intricate dependencies, aiming to automate this complex decision-making process.

A substantial body of work has investigated deep learning approaches for EHR-based medication recommendation, aiming to infer personalized drug combinations from patients' structured clinical codes and longitudinal histories. While early *instance-based methods* [19, 82] focused on extracting features from single visits, the field has largely shifted towards longitudinal modeling. Specifically, *longitudinal approaches* [3, 41] primarily utilize Recurrent Neural Networks (RNNs) [21, 77] to model long-term disease progression and treatment trajectories by aggregating temporal dependencies from patients' historical records. Concurrently, *graph-based models* have emerged to reinforce medication safety and structural understanding. These methods employ Graph Neural Networks (GNNs) [4, 55] to incorporate external knowledge, ranging from drug-drug interaction (DDI) graphs [77, 78, 88] to molecular structural information, thereby explicitly capturing the complex relational associations among diagnoses, procedures, and medications. More recently, *Large Language Models* (LLMs) [15, 42, 74, 89] have introduced a new generative paradigm, leveraging their exceptional natural language understanding and extensive medical knowledge for more comprehensive patient modeling.

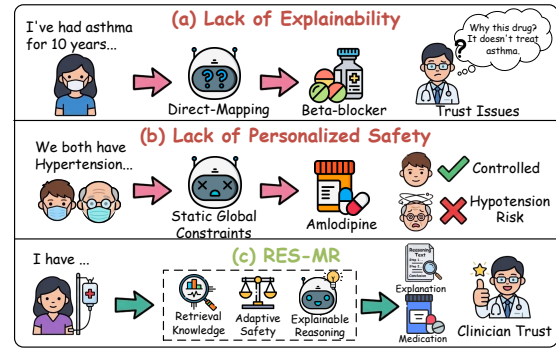


Figure 1: Examples illustrating two limitations of existing methods. (a) Most RNN/GNN/LLM methods follow a “direct-mapping” paradigm and usually output drugs (e.g., a beta-blocker for asthma) without providing an explicit rationale, leading to trust issues. (b) Safety-oriented methods use static global constraints, imposing a one-size-fits-all risk tolerance and thus risking vulnerable patients or over-restricting robust ones. (c) RES-MR leverages retrieval-augmented reasoning and personalized boundaries to generate verifiable explanations and safe medication recommendations.

Despite these advances, existing methods remain limited by opaque decision processes and reliance on fixed, patient-agnostic safety constraints without considering specific patient conditions. This oversimplified formulation fails to reflect the complexity of real-world clinical practice, as illustrated in Figure 1, leaving two critical challenges (CHs) unaddressed:

• **CH1: How to generate a *transparent, step-by-step reasoning path* that doctors can verify easily?** As illustrated in Figure 1 (a), the prevalent “direct mapping” paradigm, seen in both discriminative GNNs [45] and generative LLM-based methods [74, 85], bypasses the explicit reasoning path that should connect symptoms, pathology, and treatment rationale. Although some efforts have attempted to mitigate this opacity via attention mechanisms or aspect-based alignment [45], they primarily provide post-hoc feature attribution, rather than generating explicit diagnostic logic. They highlight which clinical features contributed to the prediction, but fail to articulate how the conclusion was logically deduced. Such a lack of transparency leads to the dilemma of unverifiable errors. For instance, if a model erroneously recommends a beta-blocker to an asthmatic patient, an opaque system cannot clarify whether the error stems from a misinterpretation of the disease status or neglect of pharmacological contraindications. Without a verifiable reasoning path, clinicians often struggle to validate or trust the system’s recommendations in high-stakes scenarios.

• **CH2: How to balance risks and benefits under an *adaptive, patient-specific risk boundary* informed by individual clinical conditions?** As shown in Figure 1 (b), most methods [19, 21, 70] enforce safety through static, global constraints such as DDI penalties in the loss function [77] or molecular-structure-informed representations [8, 83], without accounting for heterogeneity in patient-specific risk tolerance. In practice, acceptable risk is highly dependent on individual clinical status. Consider hypertension treatment as an example:

a physiologically robust young patient may tolerate a more aggressive strategy to maximize long-term control, whereas a frail older patient with multiple comorbidities requires conservative management to avoid adverse events (e.g., hypotension), even at the cost of less stringent blood pressure control. Due to the overlook of such patient-specific tolerance differences, existing models struggle to optimize utility under the patient-specific risk boundary. As a result, their provided recommendations are either overly conservative for robust patients or potentially risky for vulnerable ones.

To bridge these gaps, we propose **RES-MR**, a novel Risk-aware Reasoning framework for **Explainable and Safe Medication Recommendation** using LLMs, as shown in Figure 2. RES-MR adopts a two-stage training paradigm integrating Supervised Fine-Tuning (SFT) [23] and Reinforcement Learning (RL) [16, 87] to respectively address the aforementioned two challenges. **Stage 1: Explainable Clinical Reasoning Distillation (ECRD)**. To address CH1 and provide transparent, verifiable decision paths, we apply SFT to endow the LLM with generative reasoning capabilities. First, to ensure the reasoning is grounded in clinical facts rather than hallucinations, we introduce a knowledge graph retrieval module to extract patient-specific evidence. Then, we employ a reasoning distillation strategy, fine-tuning the LLM on high-quality reasoning trajectories to explicitly articulate logical chains as explanations (an example is shown in Figure 2). By connecting symptoms to pharmacological mechanisms based on retrieved evidence, this stage effectively ensures the decision process is explainable and clinically verifiable. **Stage 2: Risk-Aware Policy Optimization (RAPO)**. To address CH2 and achieve personalized risk-benefit trade-offs, we align the model with patient-specific safety constraints via RL. The core idea is to guide the model using adaptive reward shaping to explore effective treatments without exceeding patient’s risk tolerance. To quantify this tolerance, we first design a dual-factor disentanglement module to decompose drug representations into efficacy and risk factors. These factors are then aggregated into patient prototypes, representing desired medication and avoided medication. Leveraging the patient’s historical records and constructed prototypes, we can calibrate a personalized risk boundary. Furthermore, we present a reward clipping strategy to prevent potentially imbalanced rewards. Based on this two-stage training process, RES-MR generates explicit reasoning trajectories as explanations while providing personalized safe medication recommendations.

In summary, the main contributions of this paper are as follows:

- We propose RES-MR, a novel framework for explainable and safe medication recommendation using LLMs. It integrates reasoning generation and safety alignment into a two-stage paradigm, effectively addressing the challenges of opaque decision processes and patient-agnostic safety constraints.
- In Stage 1, we propose explainable clinical reasoning distillation to ensure transparent decision-making. By integrating the knowledge graph retrieval module with a reasoning distillation strategy, we endow the model with the capability of generating explainable reasoning trajectories grounded in clinical evidence.
- In Stage 2, we devise risk-aware policy optimization to achieve personalized risk-benefit trade-offs. We introduce a dual-factor disentanglement module and patient prototypes to quantify personalized risk boundaries, which guide an RL policy via adaptive

reward shaping and clipping to ensure safe exploration tailored to individual patient tolerance.

- Extensive experiments on benchmark datasets demonstrate that RES-MR consistently outperforms state-of-the-art baselines across recommendation accuracy, safety, and explainability metrics, validating its potential for reliable clinical decision support.

2 Related Works

Medication Recommendation. Based on the utilization of clinical data, existing studies can be categorized into instance-based and longitudinal methods. Instance-based methods [19, 82], such as LEAP [82], primarily rely on the current visit information to encode diagnoses through multi-instance multi-label learning, often overlooking the long-term progression of diseases. In contrast, longitudinal methods incorporate patient history to model health trajectories [30, 31, 54, 63, 67, 80, 84]. For instance, COGNet [71] introduces a copy-or-predict mechanism to reuse historical medications, while ARMR [66] utilizes segmented temporal learning components to capture evolving patient statuses. Regarding transparency, ExpDrug [45] attempts to map patient features into explainable aspect features to improve model interpretability. Recently, Large Language Models (LLMs) have introduced a generative paradigm to this field [15, 42, 74, 85]. LAMO [85] integrates structured codes with unstructured clinical text for the first time.

Safe Medication Recommendation. Mitigating adverse drug events and ensuring safe medication recommendations are paramount objectives in clinical applications [21, 56]. Early approaches like SafeDrug [77] employ dual molecular graph encoders to capture structural drug information and minimize DDI losses, while MoleRec [78] enhances safety by modeling the relationships between critical molecular substructures and patient health conditions. ARC [7] improves safety from a reliability perspective by calibrating confidence for medication recommendations at both the individual-drug and set levels. With the advent of LLMs, recent works have attempted to integrate safety constraints into generative frameworks. For example, KELLM [74] incorporates safety constraints via LoRA fine-tuning with external knowledge, while FLAME [15] formulates recommendation as a sequential decision process, utilizing step-wise Group Relative Policy Optimization for fine-grained drug list generation. Despite these advancements, existing methods lack verifiable reasoning trajectories and overlook patient-specific risk tolerance, hindering the adaptive balance between efficacy and safety across diverse patients.

Reasoning-enhanced LLM-Based Recommendation. Recent advancements have spurred interest in leveraging the reasoning capabilities of Large Language Models (LLMs) to move beyond simple rating prediction toward reasoning-enhanced recommendation [2, 11, 34, 39]. While in-context learning strategies like Chain-of-Thought (CoT) have been explored, general-purpose LLMs often lack domain-specific optimization without fine-tuning. Consequently, research has shifted toward fine-tuning paradigms to internalize reasoning patterns [17, 35, 40, 57]. EXP3RT [29] proposes a distillation framework that constructs profiles from reviews to jointly generate rationales and ratings in a single step. Reason4Rec [17] introduces a deliberative, multi-stage framework supervised by feedback to enable finer control over reasoning paths.

However, these general-domain methods prioritize user preference alignment, failing to address the complexity of clinical practice where accurate decisions require jointly considering therapeutic benefits, drug-drug interactions, and contraindications.

3 Problem Formulation

Electronic Health Records (EHRs). Let $\mathcal{D}$, $\mathcal{P}$, and $\mathcal{M}$ denote the sets of diagnoses, procedures, and medications, and $|\mathcal{D}|$, $|\mathcal{P}|$, and $|\mathcal{M}|$ represent their sizes. A patient’s EHR is composed of a temporal sequence of visits $\mathcal{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T]$, where the t -th visit is a tuple $\mathbf{v}_t = (C_t, \mathbf{n}_t, \mathbf{m}_t)$. Here, $C_t = [\mathbf{d}_t, \mathbf{p}_t, \mathbf{a}_t]$ denotes the structured clinical context with the diagnosis set $\mathbf{d}_t \subseteq \mathcal{D}$, the procedure set $\mathbf{p}_t \subseteq \mathcal{P}$, and patient’s demographic attributes $\mathbf{a}_t$. In contrast, $\mathbf{n}_t$ is the unstructured clinical note, and $\mathbf{m}_t \subseteq \mathcal{M}$ denotes the medication set prescribed in the visit. In our framework, the structured context C_t , the clinical note $\mathbf{n}_t$, and the medications in $\mathbf{m}_t$ are all described in natural language.

Medical Knowledge Graph. To support risk-aware reasoning, we utilize an external medical knowledge graph $\mathcal{G}_{kg} = (\mathcal{E}, \mathcal{R}, \mathcal{T}_{kg})$. Here, $\mathcal{E}$ represents the set of medical entities, and $\mathcal{R}$ denotes various clinical relations (e.g., *Indication* and *Contraindication*). $\mathcal{T}_{kg}$ consists of triplets (h, r, t) describing the connections between entities, where $h, t \in \mathcal{E}$ denote the head and tail entities, and $r \in \mathcal{R}$ denotes the relation between them.

Task Definition. Consider a patient with a sequence of visits $\mathcal{V}$. For the t -th visit, given the historical visits $\mathbf{v}_{1:t-1} = [\mathbf{v}_1, \dots, \mathbf{v}_{t-1}]$, the current visit context $(C_t, \mathbf{n}_t)$, and the external knowledge $\mathcal{G}_{kg}$, our goal is to train an LLM π_θ , such that it can explicitly generate: (1) a reasoning trajectory $\mathcal{T}_t^{\text{reason}}$ that explains the clinical decision process; and (2) a safe medication set $\hat{\mathbf{m}}_t$ for the current visit.

4 Methodology

As illustrated in Figure 2, RES-MR aims to generate explainable and safe medication recommendations by leveraging the exceptional reasoning capabilities of LLMs. It adopts a two-stage training paradigm. The first stage, **Explainable Clinical Reasoning Distillation (ECRD)**, focuses on addressing the opacity of decision-making. By integrating knowledge graph retrieval with a reasoning distillation strategy, it teaches the model via SFT to articulate step-by-step reasoning trajectories grounded in clinical evidence. Building upon this foundation, the second stage, **Risk-Aware Policy Optimization (RAPO)**, is devised to enforce patient-specific risk-benefit trade-offs through a reward-guided RL process. To quantify personalized risk boundaries, it first decomposes drug representations into efficacy and risk factors and aggregates them into prototypes that capture desired and avoided medications. Using these prototypes together with the patient’s historical records, we calibrate a personalized risk boundary. Finally, we optimize the policy via adaptive reward shaping and clipping, guiding the model to explore effective treatments while strictly suppressing boundary violations.

4.1 Explainable Clinical Reasoning Distillation

To enable the generation of verifiable reasoning paths for the LLM π_θ , this stage comprises two steps: (1) knowledge graph retrieval to construct evidence that grounds the reasoning process in objective medical facts; and (2) supervised fine-tuning on synthesized

reasoning trajectories to guide the model to explicitly articulate diagnostic logic as explanations.

4.1.1 Knowledge Graph Retrieval. For a given patient at the current visit, we perform both *node* and *semantic* retrieval on the Medical Knowledge Graph to construct the patient-specific knowledge context [47, 61, 75]. First, to capture explicit structural relations, we employ *node retrieval* by querying the KG with the current structured context C_t (e.g., diagnoses and procedures) and collecting one-hop neighboring triples of these entities to form $\mathcal{G}_{\text{node}}$. Complementarily, to uncover implicit semantic associations, we perform *semantic retrieval*. We embed the current note $\mathbf{n}_t$ into a dense vector space and retrieve the most relevant KG contexts via cosine similarity search, yielding the semantic knowledge context $\mathcal{G}_{\text{sem}}$. Finally, we merge the two streams into a unified knowledge context $\mathcal{G}_{\text{ret}} = \mathcal{G}_{\text{node}} \cup \mathcal{G}_{\text{sem}}$, which not only provides patient-specific evidence for reasoning but also serves as the knowledge foundation for patient prototype construction.

4.1.2 Supervised Fine-tuning. In practice, clinicians formulate prescriptions following a strong rationale from the patient’s specific diagnoses. To endow the model with such diagnostic reasoning capabilities, we leverage GPT-5 to synthesize high-quality reasoning trajectories $\mathcal{T}_t^{\text{reason}}$ [20, 28] for model fine-tuning. Specifically, GPT-5 is prompted with a comprehensive patient context, including the structured context C_t and note $\mathbf{n}_t$, the retrieved knowledge $\mathcal{G}_{\text{ret}}$ for grounding, and the ground-truth medication $\mathbf{m}_t$. The resulting trajectory provides step-by-step, diagnosis-aware explanations that connect patient evidence and symptoms to pharmacological mechanisms and medication actions. A clear example is shown in Figure 2. We fine-tune the LLM π_θ using the synthesized dataset $\mathcal{D}$ with low-rank adaptation [23]. Each instance includes the patient history $\mathbf{v}_{1:t-1}$, current visit inputs $(C_t, \mathbf{n}_t)$, and retrieved knowledge $\mathcal{G}_{\text{ret}}$. We minimize the negative log-likelihood of generating y :

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\sum_{j=1}^{|y|} \log \pi_\theta(y_j \mid x, y_{<j}) \right], \quad (1)$$

where we define the input $x = (\mathbf{v}_{1:t-1}, C_t, \mathbf{n}_t, \mathcal{G}_{\text{ret}})$ and y represents the concatenated token sequence of the reasoning path $\mathcal{T}_t^{\text{reason}}$ and the final result $\mathbf{m}_t$. By learning this structured generation pattern, the LLM acquires diagnosis-aware reasoning behaviors and a strong, format-stable initialization for the RAPO stage.

4.2 Risk-Aware Policy Optimization

While SFT establishes fundamental clinical reasoning, its supervision objective overlooks the efficacy-safety trade-off. Consequently, the model struggles to satisfy personalized risk constraints while maintaining accuracy. Therefore, we propose Risk-Aware Policy Optimization (RAPO), which incorporates the trade-off via specially devised reinforcement learning [58, 79]. The core objective is to guide the model to explore more effective medication combinations while adhering to personalized risk boundaries. To achieve this, we first pretrain a dual-factor disentanglement module to decompose drug representations into efficacy and risk factors, making both signals explicitly quantifiable for reward computation. Based on retrieved knowledge, we then construct patient prototypes that

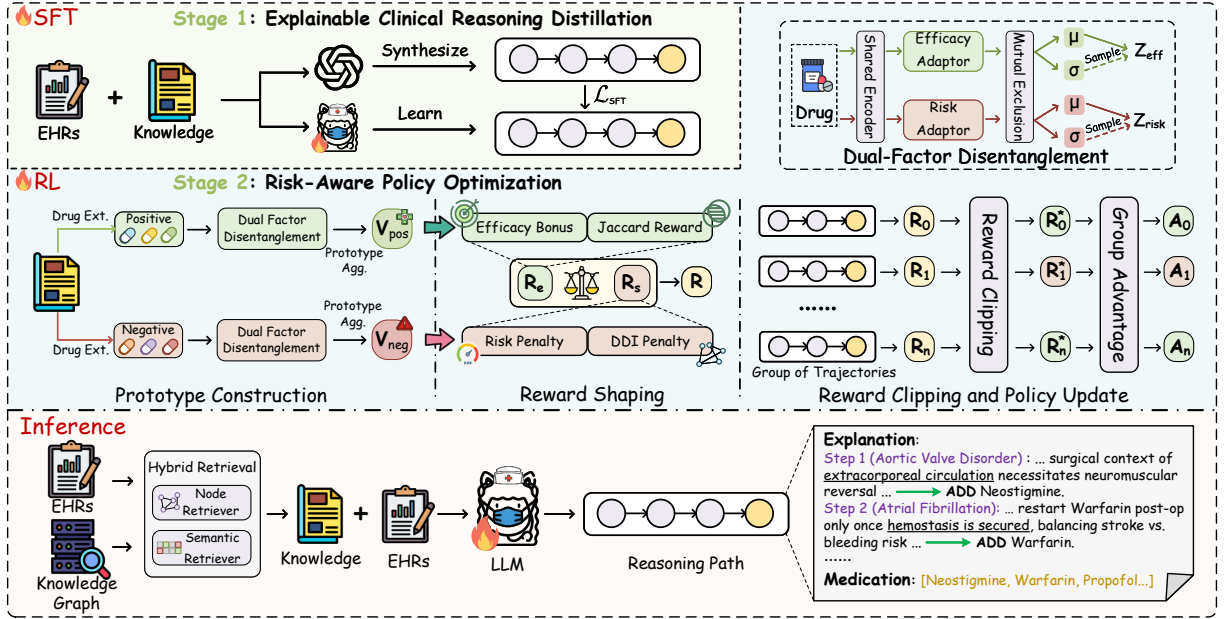


Figure 2: Overall framework and training/inference pipeline of RES-MR. Stage 1: Given patient EHRs, hybrid retrieval retrieves patient-specific evidence from the knowledge graph; a teacher LLM synthesizes reasoning trajectories to perform Supervise Fine-tuning (SFT) of the LLM, endowing it with strong reasoning capabilities to provide explicit explanations. Stage 2: We further optimize the LLM with Reinforcement Learning (RL). A pretrained dual-factor disentanglement extracts drug efficacy and risk factors to construct patient-specific prototypes, which guide the policy update during RL training via reward shaping and clipping to ensure safety. Inference: The fully trained LLM uses the hybrid retrieval to augment generation with retrieved knowledge and outputs verifiable explanations and safe medication recommendations.

summarize desired and avoided medications. Using these prototypes together with the patient’s medication history, we calibrate a patient-specific risk boundary. In this stage, these signals are used for reward shaping to balance efficacy and safety. We further apply reward clipping to prevent potentially imbalanced rewards.

4.2.1 Dual-Factor Disentanglement. To make efficacy and risk signals quantifiable for reward measurement, we introduce a *Dual-Factor Disentanglement* module and pretrain it before RL [44]. For each drug m , we obtain its semantic embedding $\mathbf{e}_m$ from an LLM encoder. We then encode $\mathbf{e}_m$ with a shared encoder and two factor-specific adaptors, i.e., $\mathbf{h} = f_{\text{share}}(\mathbf{e}_m)$, $\mathbf{h}_{\text{eff}} = f_{\text{eff}}(\mathbf{h})$, and $\mathbf{h}_{\text{risk}} = f_{\text{risk}}(\mathbf{h})$. To suppress shared information across factors, we introduce a *Mutual Exclusion Block* [12] prior to the variational projection:

$$\mathbf{h}'_{\text{eff}} = \mathbf{h}_{\text{eff}} - \alpha \cdot \sigma(W_g[\mathbf{h}_{\text{eff}}; \mathbf{h}_{\text{risk}}]) \odot \text{MLP}(\mathbf{h}_{\text{risk}}), \quad (2)$$

where σ is the sigmoid function, W_g is a learnable weight matrix, and α is a learnable parameter. A symmetric operation is also applied to $\mathbf{h}'_{\text{risk}}$. For $k \in \{\text{eff}, \text{risk}\}$, we project $\mathbf{h}'_k$ to (μ_k, σ_k) and adopt the reparameterization trick [32]: $\mathbf{z}_k = \mu_k + \epsilon \odot \sigma_k$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We align each factor with clinical semantics using KG relations. Specifically, the positive concept set T_k^+ is constructed from $\mathcal{T}_{kg}$: for efficacy ($k = \text{eff}$), positives are clinical concepts linked by *Indication*; for risk ($k = \text{risk}$), positives are those linked by *Contraindication*. Negatives T_k^- are randomly sampled from other concepts. For each

factor k , we pull $\mathbf{z}_k$ close to T_k^+ and push it away from T_k^- :

$$\mathcal{L}_{\text{align}}^{(k)} = - \sum_{t \in T_k^+} \log \sigma(s(\mathbf{z}_k, t)) - \sum_{t' \in T_k^-} \log(1 - \sigma(s(\mathbf{z}_k, t'))), \quad (3)$$

where $s(\cdot, \cdot)$ denotes cosine similarity. The overall alignment loss is $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{align}}^{(\text{eff})} + \mathcal{L}_{\text{align}}^{(\text{risk})}$. To promote independence and robustness, we impose an orthogonality constraint and the KL regularizer [24]:

$$\mathcal{L}_{\text{orth}} = \|\text{Cov}(\mu_{\text{eff}}, \mu_{\text{risk}})\|_F^2, \quad (4)$$

$$\mathcal{L}_{\text{KL}} = \sum_{k \in \{\text{eff}, \text{risk}\}} D_{\text{KL}}(q_\phi(\mathbf{z}_k | \mathbf{e}_m) \| \mathcal{N}(\mathbf{0}, \mathbf{I})), \quad (5)$$

where Cov is covariance matrix and D_{KL} is KL divergence. We parameterize variational posterior: $q_\phi(\mathbf{z}_k | \mathbf{e}_m) = \mathcal{N}(\mu_k, \text{diag}(\sigma_k^2))$. The final objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \beta_{\text{KL}} \mathcal{L}_{\text{KL}}$. After convergence, we use the frozen μ_{eff} and μ_{risk} as deterministic drug factors for the following optimization.

4.2.2 Prototype Construction. Given a patient and the current visit, this module constructs patient-specific prototypes as reference points [49] for reward computation in RAPO. Leveraging the pretrained disentangled factors and retrieved knowledge context $\mathcal{G}_{\text{ret}}$, we derive these prototypes to capture the desired therapeutic features and the risk features to avoid. Specifically, based on medical relations in the Knowledge Graph (e.g., indications and contraindications), we categorize drug entities from $\mathcal{G}_{\text{ret}}$ into an indication set $\mathcal{M}_{\text{pos}}$ and a contraindication set $\mathcal{M}_{\text{neg}}$. To ensure valid prototype construction, we require both sets to be non-empty for each patient. Subsequently, utilizing the frozen encoder, we aggregate

the corresponding disentangled factors of the drugs within each set. The *positive prototype* $\mathbf{v}_{\text{pos}}$ characterizes the desired therapeutic direction by aggregating the efficacy factors of the indicated drugs:

$$\mathbf{v}_{\text{pos}} = \frac{1}{|\mathcal{M}_{\text{pos}}|} \sum_{m \in \mathcal{M}_{\text{pos}}} \boldsymbol{\mu}_{\text{eff}}(m). \quad (6)$$

Conversely, the *negative prototype* $\mathbf{v}_{\text{neg}}$ captures the high-risk profiles that must be strictly avoided, obtained by aggregating the risk factors of the contraindicated drugs:

$$\mathbf{v}_{\text{neg}} = \frac{1}{|\mathcal{M}_{\text{neg}}|} \sum_{m \in \mathcal{M}_{\text{neg}}} \boldsymbol{\mu}_{\text{risk}}(m). \quad (7)$$

4.2.3 Reward Shaping. To guide the model toward optimizing the trade-off between therapeutic efficacy and personalized safety, we design a composite reward function that combines efficacy rewards with safety costs under a patient-specific risk boundary. Prior to quantifying safety costs, we first calibrate a risk boundary B to define the maximum risk level a patient can tolerate. Using historical medications $\mathcal{M}_{\text{hist}} = \bigcup_{i < t} \mathbf{m}_i$ as an empirical proxy [48, 50] of patient-specific risk exposure under prior clinical supervision, we define B as the δ -quantile of risk similarities between historical drugs and the negative prototype $\mathbf{v}_{\text{neg}}$:

$$B = \inf \left\{ b \in \mathbb{R} \mid \frac{1}{|\mathcal{M}_{\text{hist}}|} \sum_{m \in \mathcal{M}_{\text{hist}}} \mathbb{I}(s(\boldsymbol{\mu}_{\text{risk}}(m), \mathbf{v}_{\text{neg}}) \leq b) \geq \delta \right\}, \quad (8)$$

where $s(\cdot, \cdot)$ denotes cosine similarity and $\delta \in [0, 1]$ controls the strictness. A larger B implies higher tolerance, allowing broader exploration, whereas a smaller B imposes tighter constraints for vulnerable patients. The final reward R is synthesized by subtracting a weighted safety cost from the efficacy reward:

$$R = (1 - \rho_t)R_e - \rho_t R_s. \quad (9)$$

The *efficacy reward* R_e encourages clinically accurate prescriptions by combining the Jaccard similarity with the ground truth $\mathbf{m}^*$ and an efficacy bonus based on the positive prototype $\mathbf{v}_{\text{pos}}$:

$$R_e = \lambda_j \frac{|\hat{\mathbf{m}} \cap \mathbf{m}^*|}{|\hat{\mathbf{m}} \cup \mathbf{m}^*|} + \lambda_e \frac{1}{|\hat{\mathbf{m}}|} \sum_{d \in \hat{\mathbf{m}}} \max(0, s(\boldsymbol{\mu}_{\text{eff}}(d), \mathbf{v}_{\text{pos}})). \quad (10)$$

This efficacy bonus ensures the model receives positive feedback for functionally valid drugs aligned with the desired therapeutic direction, even if they differ from the exact labels. Conversely, the *safety cost* R_s penalizes iatrogenic risks, aggregating Drug-Drug Interaction (DDI) rates and violations of the calibrated risk boundary:

$$R_s = \lambda_d \text{DDI}(\hat{\mathbf{m}}) + \lambda_r \frac{1}{|\hat{\mathbf{m}}|} \sum_{d \in \hat{\mathbf{m}}} \max(0, s(\boldsymbol{\mu}_{\text{risk}}(d), \mathbf{v}_{\text{neg}}) - B). \quad (11)$$

The second term penalizes only the amount by which a drug's risk similarity to $\mathbf{v}_{\text{neg}}$ exceeds the boundary B . Crucially, any format violation or hallucinated drug is assigned a fixed penalty ($R = -100$) and treated as an invalid sample, bypassing the computation of R_e and R_s . Finally, to prevent early convergence to overly conservative behaviors, we introduce an adaptive gating coefficient ρ_t derived from the moving average Jaccard performance J_t :

$$\rho_t = \sigma(\bar{J}_t - \tau_t), \quad (12)$$

where τ_t linearly increases during training. This mechanism ensures that strict safety constraints intervene only after the model acquires sufficient clinical reasoning accuracy.

4.2.4 Reward Clipping and Policy Update. With the established reward shaping, we may still encounter a critical risk: high efficacy rewards could potentially overshadow safety costs, encouraging the model to trade safety for performance (i.e., reward hacking [6, 68]). To address this, we introduce a reward clipping strategy that suppresses “high-reward but unsafe” trajectories upon boundary violations, while allowing a higher upper bound for safe trajectories, thereby encouraging safe exploration and conservative updates, respectively [9, 16]. Specifically, we compute the difference between the personalized risk boundary B and the maximum risk similarity among the generated medications, and use it to adaptively clip the raw reward R :

$$\Delta_t = B - \max_{d \in \hat{\mathbf{m}}} s(\boldsymbol{\mu}_{\text{risk}}(d), \mathbf{v}_{\text{neg}}). \quad (13)$$

A positive Δ_t indicates the medication set is within the safe region, while a non-positive Δ_t signals a boundary violation. Based on the sign of Δ_t , we obtain the final reward R_{total} by clipping the raw shaped reward R , where the clipping bounds are computed from the group-wise mean μ_R and standard deviation σ_R over the G sampled trajectories for the same input context:

$$R_{\text{total}} = \begin{cases} \text{clip}(R, \mu_R - \sigma_R, \mu_R + 2\sigma_R) & \text{if } \Delta_t > 0 \\ \text{clip}(R, \mu_R - 2\sigma_R, \mu_R + \sigma_R) & \text{if } \Delta_t \leq 0 \end{cases}. \quad (14)$$

Here $\text{clip}(x, a, b) = \min(\max(x, a), b)$ truncates the reward into $[a, b]$, preventing extreme reward values from dominating policy updates and stabilizing optimization. When $\Delta_t > 0$, we relax the upper bound to allow higher rewards for safe trajectories. When $\Delta_t \leq 0$, we tighten the upper bound and relax the lower bound to suppress unsafe high-reward trajectories, mitigating reward hacking and inducing conservative updates.

Finally, to ensure stable convergence under these dynamic constraints, we employ Group Sequence Policy Optimization [87]. Concretely, we sample an input context $x \sim \mathcal{D}$. For each x , we draw G candidate sequences $\{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | x)$, where each y_i is a complete generation containing the reasoning trajectory and the medication set. We maximize the following objective:

$$\mathcal{J}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i \right) \right], \quad (15)$$

where $\hat{A}_i$ is the standardized advantage derived from R_{total} . The importance ratio $s_i(\theta)$ is defined based on sequence likelihood:

$$s_i(\theta) = \left(\frac{\pi_{\theta}(y_i | x)}{\pi_{\theta_{\text{old}}}(y_i | x)} \right)^{\frac{1}{|y_i|}}. \quad (16)$$

This length normalization keeps the importance ratio comparable across varying response lengths, which helps stabilize training.

4.3 Inference

During inference, we perform knowledge graph retrieval to obtain $\mathcal{G}_{\text{ret}}$ for the current visit. We form the input context as $x =$

Table 1: Statistics on MIMIC-III, MIMIC-IV, and eICU Datasets

Items	MIMIC-III	MIMIC-IV	eICU
# of visits / # of patients	12,433 / 4,785	156,315 / 46,811	18,636 / 8,544
dis. / proc. / med. space size	4,316 / 1,330 / 593	19,812 / 10,859 / 743	1,562 / 937 / 152
avg. / max # of visits	5.68 / 29	11.07 / 70	2.18 / 9
avg. / max # of dis. per visit	14.68 / 39	13.45 / 56	4.79 / 76
avg. / max # of proc. per visit	4.69 / 32	2.75 / 34	8.03 / 118
avg. / max # of med. per visit	14.78 / 67	14.24 / 128	10.35 / 51

$(\mathbf{v}_{1:t-1}, \mathbf{C}_t, \mathbf{n}_t, \mathcal{G}_{\text{ret}})$ and autoregressively generate a serialized output sequence:

$$\hat{y} \sim \pi_{\theta}(\cdot | x). \quad (17)$$

The generated $\hat{y}$ for visit t consists of an explicit reasoning trajectory $\hat{\mathcal{T}}_t^{\text{reason}}$ followed by the recommended medication set $\hat{\mathbf{m}}_t$.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on MIMIC-III [27], MIMIC-IV [26], and eICU [51]. MIMIC-III and MIMIC-IV are large-scale databases comprising deidentified critical care data from the Beth Israel Deaconess Medical Center. eICU is a multi-center ICU EHR database with deidentified patient records across hospitals. To better reflect real-world clinical practice, we retain the complete set of diagnoses, procedures, and medications without truncation. Additionally, for the MIMIC-III and MIMIC-IV datasets, we only include admissions that have available clinical notes. Regarding text processing, we follow the common practice [15, 85] and extract specific segments from these notes, including history of present illness, past medical history, allergies, and medications on admission. For data standardization, we unify the diagnosis and procedure codes by mapping them to the ICD-9 standard [22], utilizing their descriptions as model inputs. Similarly, medications are mapped to DrugBank IDs [65], with their corresponding generic names to unify diverse formulations. Furthermore, following established protocols [37, 77], we construct the drug-drug interaction matrix $\mathcal{M}_{ddi}$ based on interaction pairs extracted from the DrugBank database [65]. We split all three datasets into training, validation, and test sets at the patient level with a 4:1:1 ratio. Retaining the full medication set enlarges the label space, making results not directly comparable to prior studies with different preprocessing. The detailed statistics of the processed datasets are summarized in Table 1. **Metrics.** We evaluate our method from three perspectives: recommendation accuracy, medication safety, and reasoning interpretability. We adopt the commonly-used accuracy metrics, i.e., **Jaccard Similarity (Jaccard)** [1] and **F1-score (F1)** [53], to measure the overlap between the predicted set $\hat{\mathbf{m}}_i$ and the ground-truth set $\mathbf{m}_i$ at the last visit for each patient i in the test set $\mathcal{U}$:

$$\text{Jaccard} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \frac{|\mathbf{m}_i \cap \hat{\mathbf{m}}_i|}{|\mathbf{m}_i \cup \hat{\mathbf{m}}_i|}, \text{F1} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \frac{2 \cdot \text{Pre}_i \cdot \text{Rec}_i}{\text{Pre}_i + \text{Rec}_i}, \quad (18)$$

where $\text{Pre}_i = \frac{|\hat{\mathbf{m}}_i \cap \mathbf{m}_i|}{|\hat{\mathbf{m}}_i|}$ and $\text{Rec}_i = \frac{|\hat{\mathbf{m}}_i \cap \mathbf{m}_i|}{|\mathbf{m}_i|}$. Safety is assessed via the **DDI rate** [55] and our proposed **Risk Intrusion Score (RIS)**. The DDI rate calculates the proportion of medication pairs in $\hat{\mathbf{m}}_i$ that

induce interactions based on the adjacency matrix $\mathbf{M}_{ddi}$:

$$\text{DDI} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \frac{\sum_{j=1}^{|\hat{\mathbf{m}}_i|} \sum_{k=j+1}^{|\hat{\mathbf{m}}_i|} \mathbb{I}(\mathbf{M}_{ddi}[\hat{\mathbf{m}}_i^{(j)}, \hat{\mathbf{m}}_i^{(k)}] = 1)}{\sum_{j=1}^{|\hat{\mathbf{m}}_i|} \sum_{k=j+1}^{|\hat{\mathbf{m}}_i|} 1}, \quad (19)$$

where $\mathbb{I}(\cdot)$ is the indicator function. To empirically assess patient-specific risk tolerance, which is not directly observable from EHRs, we define **RIS** as a proxy metric [50] that measures the extent to which predicted drugs violate each patient’s risk boundary B_i :

$$\text{RIS} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \frac{1}{|\hat{\mathbf{m}}_i|} \sum_{d \in \hat{\mathbf{m}}_i} \frac{\max(0, s(\boldsymbol{\mu}_{\text{risk}}(d), \mathbf{v}_{\text{neg}}) - B_i)}{1 - B_i}. \quad (20)$$

Grounded in external medical knowledge, RIS leverages frozen disentangled factors as a safety signal complementary to DDI, capturing patient-specific risks. To evaluate the quality of explanations provided by reasoning trajectory, we follow previous work [17, 38, 46] and adopt BLEURT. Specifically, **BLEURT** measures the semantic alignment between the generated reasoning paths and reference explanations [52]. Meanwhile, we introduce two composite metrics based on harmonic means to measure the pair-wise trade-offs between accuracy, safety, and explainability: F_{AS} (Accuracy-Safety) and F_{AE} (Accuracy-Explainability):

$$F_{AS} = \frac{2 \cdot S_{\text{Acc}} \cdot S_{\text{Safe}}}{S_{\text{Acc}} + S_{\text{Safe}}}, \quad F_{AE} = \frac{2 \cdot S_{\text{Acc}} \cdot \text{BLEURT}}{S_{\text{Acc}} + \text{BLEURT}}, \quad (21)$$

where $S_{\text{Acc}} = \frac{\text{Jaccard} + \text{F1}}{2}$ and $S_{\text{Safe}} = \frac{(1 - \text{DDI}) + (1 - \text{RIS})}{2}$.

Baseline Methods. To comprehensively evaluate RES-MR, we select representative and state-of-the-art baselines and group them into three categories: (1) **General medication recommendation methods:** LEAP [82], COGNet [71], RAREMed [86], ARMR [66], LAMO [85]; (2) **Safety methods:** SafeDrug [77], MoleRec [78], ARC [7], KELLM [74], FLAME [15]; and (3) **Explainable methods:** ExpDrug [45], Reason4Rec [17], EXP3RT [29]. Among them, the **LLM-based methods** include KELLM, LAMO, FLAME, Reason4Rec, and EXP3RT, while the remaining baselines are conventional methods. To ensure a rigorous and fair comparison, we employ MICRON [76] as the backbone of ARC, following the best practice reported in the original paper [7]. For LLM-based methods, we use Llama-3.1-Aloe-Beta-8B [18] as backbone to control for foundation model capacity. Moreover, we transfer Reason4Rec and EXP3RT to our medication recommendation setting.

Implementation Details. All experiments are conducted on a cluster equipped with 8 NVIDIA A100-SXM4 GPUs, Intel Xeon Platinum 8370C CPUs, and 512GB RAM. For the **Dual-Factor Disentanglement module**, the drug encoder is configured with latent/shared dimensions of 128/512 and pretrained for 100 epochs using Adam with a learning rate of 10^{-4} . After pretraining, we freeze the encoder as a fixed feature extractor. We search λ_{orth} in $[0.05, 0.50]$ with step size 0.05 and β_{KL} in $\{1, 3, 5, 7, 9\} \times 10^{-4}$. For **Knowledge Graph Retrieval**, we leverage PrimeKG [5], performing 1-hop neighbor expansion via Neo4j and employing bge-large-en-v1.5 [73] to retrieve the top-5 entities relevant to clinical notes. In **Stage 1**, we employ GPT-5 as the teacher model and Llama-3.1-Aloe-Beta-8B [18] as the student model. We set the context length to 8,192 and retain only the latest three visits to limit input length while preserving space for retrieved knowledge. We fine-tune the student via LoRA (Rank $r = 8$, Alpha $\alpha = 16$, Dropout 0.05). The training uses a learning

Table 2: Comparison of accuracy, safety, and explainability performance with baselines on the MIMIC-III, MIMIC-IV, and eICU datasets. $\uparrow$ and $\downarrow$ indicate higher and lower is better, respectively. The best and second-best results are highlighted in bold and underlined. * denotes statistical significance ($p < 0.05$) by a paired t -test.

Dataset	Methods	Jaccard $\uparrow$	F1 $\uparrow$	DDI $\downarrow$	RIS $\downarrow$	BLEURT ¹ $\uparrow$	F_{AS} $\uparrow$	F_{AE} $\uparrow$
MIMIC-III	LEAP	0.1934 $\pm$ 0.0023	0.3285 $\pm$ 0.0032	0.1341 $\pm$ 0.0012	0.1433 $\pm$ 0.0023	-	0.4005 $\pm$ 0.0029	-
	COGNet	0.2381 $\pm$ 0.0019	0.3628 $\pm$ 0.0032	0.1466 $\pm$ 0.0013	0.1473 $\pm$ 0.0021	-	0.4444 $\pm$ 0.0015	-
	RAREMed	0.2244 $\pm$ 0.0046	0.3414 $\pm$ 0.0012	0.1482 $\pm$ 0.0046	0.1547 $\pm$ 0.0038	-	0.4243 $\pm$ 0.0026	-
	ARMR	0.2356 $\pm$ 0.0055	0.3599 $\pm$ 0.0035	0.1536 $\pm$ 0.0025	0.1613 $\pm$ 0.0035	-	0.4400 $\pm$ 0.0040	-
	LAMO	0.2512 $\pm$ 0.0003	0.3707 $\pm$ 0.0027	0.1124 $\pm$ 0.0023	0.1427 $\pm$ 0.0027	-	0.4585 $\pm$ 0.0015	-
	SafeDrug	0.2210 $\pm$ 0.0023	0.3469 $\pm$ 0.0012	0.1054 $\pm$ 0.0014	0.1484 $\pm$ 0.0015	-	0.4285 $\pm$ 0.0012	-
	MoleRec	0.2454 $\pm$ 0.0052	0.3770 $\pm$ 0.0038	0.1044 $\pm$ 0.0007	0.1401 $\pm$ 0.0005	-	0.4595 $\pm$ 0.0034	-
	KELLM	0.2291 $\pm$ 0.0031	0.3507 $\pm$ 0.0029	0.1318 $\pm$ 0.0025	0.1403 $\pm$ 0.0040	-	0.4341 $\pm$ 0.0019	-
	ARC	0.2377 $\pm$ 0.0012	0.3610 $\pm$ 0.0038	0.1167 $\pm$ 0.0033	0.1454 $\pm$ 0.0014	-	0.4453 $\pm$ 0.0024	-
	FLAME	0.2601 $\pm$ 0.0031	0.3880 $\pm$ 0.0033	0.1291 $\pm$ 0.0009	0.1374 $\pm$ 0.0013	-	0.4717 $\pm$ 0.0026	-
	ExpDrug	0.2326 $\pm$ 0.0014	0.3610 $\pm$ 0.0022	0.1047 $\pm$ 0.0017	0.1397 $\pm$ 0.0028	-	0.4436 $\pm$ 0.0016	-
	Reason4Rec	0.1529 $\pm$ 0.0036	0.2519 $\pm$ 0.0022	0.1512 $\pm$ 0.0064	<u>0.1311 $\pm$ 0.0026</u>	0.4052 $\pm$ 0.0011	0.3276 $\pm$ 0.0022	0.2700 $\pm$ 0.0015
	EXP3RT	0.2150 $\pm$ 0.0028	0.3402 $\pm$ 0.0026	0.1589 $\pm$ 0.0014	0.1531 $\pm$ 0.0035	0.3752 $\pm$ 0.0005	0.4178 $\pm$ 0.0026	0.3191 $\pm$ 0.0015
	RES-MR-3B	0.2407 $\pm$ 0.0044	0.3681 $\pm$ 0.0017	0.1041 $\pm$ 0.0019	0.1342 $\pm$ 0.0006	<u>0.4390 $\pm$ 0.0030</u>	0.4524 $\pm$ 0.0024	<u>0.3595 $\pm$ 0.0017</u>
	RES-MR-8B	0.2803 $\pm$ 0.0046*	0.4179 $\pm$ 0.0016*	0.1031 $\pm$ 0.0006*	0.1267 $\pm$ 0.0045*	0.4538 $\pm$ 0.0026*	0.5007 $\pm$ 0.0020*	0.3946 $\pm$ 0.0017*
MIMIC-IV	LEAP	0.1761 $\pm$ 0.0055	0.3105 $\pm$ 0.0012	0.1310 $\pm$ 0.0021	0.1474 $\pm$ 0.0039	-	0.3794 $\pm$ 0.0037	-
	COGNet	0.2279 $\pm$ 0.0021	0.3598 $\pm$ 0.0008	0.1391 $\pm$ 0.0031	0.1408 $\pm$ 0.0002	-	0.4380 $\pm$ 0.0011	-
	RAREMed	0.2214 $\pm$ 0.0006	0.3447 $\pm$ 0.0007	0.1440 $\pm$ 0.0053	0.1526 $\pm$ 0.0015	-	0.4249 $\pm$ 0.0008	-
	ARMR	0.2280 $\pm$ 0.0032	0.3541 $\pm$ 0.0006	0.1499 $\pm$ 0.0010	0.1337 $\pm$ 0.0037	-	0.4347 $\pm$ 0.0022	-
	LAMO	0.2359 $\pm$ 0.0015	0.3582 $\pm$ 0.0040	0.1370 $\pm$ 0.0012	0.1410 $\pm$ 0.0041	-	0.4417 $\pm$ 0.0022	-
	SafeDrug	0.2107 $\pm$ 0.0058	0.3398 $\pm$ 0.0023	0.1104 $\pm$ 0.0034	0.1473 $\pm$ 0.0037	-	0.4183 $\pm$ 0.0041	-
	MoleRec	0.2311 $\pm$ 0.0029	0.3607 $\pm$ 0.0031	0.1084 $\pm$ 0.0050	0.1391 $\pm$ 0.0022	-	0.4424 $\pm$ 0.0033	-
	KELLM	0.2013 $\pm$ 0.0060	0.3164 $\pm$ 0.0006	0.1560 $\pm$ 0.0007	0.1401 $\pm$ 0.0016	-	0.3971 $\pm$ 0.0035	-
	ARC	0.2297 $\pm$ 0.0019	0.3586 $\pm$ 0.0043	0.1198 $\pm$ 0.0048	0.1374 $\pm$ 0.0037	-	0.4398 $\pm$ 0.0026	-
	FLAME	0.2401 $\pm$ 0.0056	0.3644 $\pm$ 0.0049	0.1210 $\pm$ 0.0024	0.1344 $\pm$ 0.0043	-	0.4489 $\pm$ 0.0036	-
	ExpDrug	0.2257 $\pm$ 0.0022	0.3516 $\pm$ 0.0027	0.1073 $\pm$ 0.0025	0.1410 $\pm$ 0.0038	-	0.4342 $\pm$ 0.0024	-
	Reason4Rec	0.1477 $\pm$ 0.0009	0.2480 $\pm$ 0.0029	0.1476 $\pm$ 0.0044	0.1370 $\pm$ 0.0018	0.4207 $\pm$ 0.0030	0.3215 $\pm$ 0.0020	0.2691 $\pm$ 0.0017
	EXP3RT	0.2055 $\pm$ 0.0019	0.3301 $\pm$ 0.0022	0.1331 $\pm$ 0.0032	0.1431 $\pm$ 0.0024	0.4077 $\pm$ 0.0027	0.4086 $\pm$ 0.0021	0.3233 $\pm$ 0.0011
	RES-MR-3B	0.2301 $\pm$ 0.0050	0.3340 $\pm$ 0.0009	0.1017 $\pm$ 0.0021	<u>0.1303 $\pm$ 0.0007</u>	<u>0.4288 $\pm$ 0.0023</u>	0.4276 $\pm$ 0.0029	<u>0.3403 $\pm$ 0.0021</u>
	RES-MR-8B	0.2525 $\pm$ 0.0012*	0.3842 $\pm$ 0.0030*	<u>0.1064 $\pm$ 0.0061</u>	0.1247 $\pm$ 0.0034*	0.4439 $\pm$ 0.0045*	0.4682 $\pm$ 0.0019*	0.3708 $\pm$ 0.0015*
eICU	LEAP	0.1831 $\pm$ 0.0031	0.3085 $\pm$ 0.0026	0.1241 $\pm$ 0.0006	0.1364 $\pm$ 0.0005	-	0.3833 $\pm$ 0.0021	-
	COGNet	0.2283 $\pm$ 0.0031	0.3580 $\pm$ 0.0012	0.1283 $\pm$ 0.0022	0.1408 $\pm$ 0.0012	-	0.4380 $\pm$ 0.0018	-
	RAREMed	0.2323 $\pm$ 0.0027	0.3601 $\pm$ 0.0027	0.1431 $\pm$ 0.0010	0.1412 $\pm$ 0.0036	-	0.4404 $\pm$ 0.0021	-
	LAMO	0.2331 $\pm$ 0.0007	0.3674 $\pm$ 0.0020	0.1173 $\pm$ 0.0012	0.1341 $\pm$ 0.0044	-	0.4470 $\pm$ 0.0010	-
	ARMR	0.2249 $\pm$ 0.0008	0.3503 $\pm$ 0.0016	0.1491 $\pm$ 0.0048	0.1401 $\pm$ 0.0024	-	0.4305 $\pm$ 0.0012	-
	SafeDrug	0.2011 $\pm$ 0.0035	0.3312 $\pm$ 0.0087	0.1104 $\pm$ 0.0034	0.1364 $\pm$ 0.0043	-	0.4083 $\pm$ 0.0056	-
	MoleRec	0.2276 $\pm$ 0.0014	0.3532 $\pm$ 0.0026	<u>0.1065 $\pm$ 0.0039</u>	0.1386 $\pm$ 0.0022	-	0.4364 $\pm$ 0.0019	-
	KELLM	0.2197 $\pm$ 0.0048	0.3482 $\pm$ 0.0017	0.1211 $\pm$ 0.0011	0.1374 $\pm$ 0.0022	-	0.4282 $\pm$ 0.0034	-
	ARC	0.2197 $\pm$ 0.0015	0.3486 $\pm$ 0.0022	0.1201 $\pm$ 0.0008	0.1433 $\pm$ 0.0007	-	0.4282 $\pm$ 0.0020	-
	FLAME	0.2397 $\pm$ 0.0043	0.3690 $\pm$ 0.0006	0.1146 $\pm$ 0.0066	0.1311 $\pm$ 0.0029	-	0.4519 $\pm$ 0.0021	-
	ExpDrug	0.2158 $\pm$ 0.0042	0.3404 $\pm$ 0.0039	0.1086 $\pm$ 0.0038	0.1324 $\pm$ 0.0027	-	0.4226 $\pm$ 0.0040	-
	Reason4Rec	0.1877 $\pm$ 0.0015	0.2871 $\pm$ 0.0013	0.1320 $\pm$ 0.0013	0.1423 $\pm$ 0.0049	<u>0.4277 $\pm$ 0.0039</u>	0.3724 $\pm$ 0.0014	0.3053 $\pm$ 0.0011
	EXP3RT	0.1981 $\pm$ 0.0044	0.3017 $\pm$ 0.0043	0.1552 $\pm$ 0.0013	0.1464 $\pm$ 0.0035	0.4119 $\pm$ 0.0049	0.3862 $\pm$ 0.0036	0.3111 $\pm$ 0.0032
	RES-MR-3B	0.2376 $\pm$ 0.0009	0.3711 $\pm$ 0.0017	0.1074 $\pm$ 0.0014	0.1291 $\pm$ 0.0046	0.4013 $\pm$ 0.0016	0.4525 $\pm$ 0.0010	0.3462 $\pm$ 0.0007
	RES-MR-8B	0.2593 $\pm$ 0.0022*	0.3908 $\pm$ 0.0039*	0.1009 $\pm$ 0.0009*	0.1247 $\pm$ 0.0016*	0.4368 $\pm$ 0.0016*	0.4758 $\pm$ 0.0021*	0.3727 $\pm$ 0.0016*

rate of 2×10^{-4} , the AdamW optimizer, and Flash Attention 2 [13]. In **Stage 2**, we initialize with Stage 1 weights, maintaining LoRA settings, and employ GSPO [87] ($G = 4$, $\epsilon = 0.2$) with a learning rate of 10^{-6} . The risk boundary strictness δ is set to 0.9, and reward weights $\lambda_{\{j,e,d,r\}}$ are tuned within $\{1, 3, 5, 7, 10\}$. The adaptive gating threshold τ_t is linearly annealed from 0.2 during training. All experiments are repeated five times with different random seeds.

5.2 Overall Performance Comparison

We evaluate RES-MR on the MIMIC-III dataset across accuracy, safety, and explainability. Meanwhile, following [10, 15], we evaluate the generalization and transferability of our RES-MR by initializing the model with SFT weights from MIMIC-III and performing only one epoch of RAPO training on MIMIC-IV and eICU datasets,

whereas baselines are fully trained on these datasets. We implement RES-MR using both Llama-3.2-3B-Instruct and Llama-3.1-AIoe-Beta-8B backbones to assess performance across different parameter scales, denoted as RES-MR-3B and RES-MR-8B.

As shown in Table 2, we report the average performance and standard deviations of each method on three datasets. From these results, we observe the following: 1) RES-MR consistently achieves the best overall performance across three datasets. In particular, RES-MR-8B outperforms all baselines, achieving a 7.7% improvement in Jaccard over the state-of-the-art FLAME on MIMIC-III. These gains are attributed to our ECRD, which fosters explicit clinical reasoning rather than mere co-occurrence patterns, and to RAPO, which

¹BLEURT and F_{AE} are only computed for methods that generate textual explanations; “-” indicates that the method does not output explanation text.

dynamically optimizes the efficacy-safety trade-off. Notably, under our transfer setting where RES-MR is initialized from MIMIC-III SFT weights and only lightly adapted on MIMIC-IV and eICU, it still maintains superior performance, demonstrating robust transferability. 2) Safety baselines can be overly conservative, making it difficult to balance accuracy and safety and leading to suboptimal performance. For instance, SafeDrug and MoleRec primarily enforce safety via DDI penalties. Consequently, these methods may prioritize interaction avoidance over therapeutic coverage, resulting in overly restricted medication combinations. In contrast, RES-MR achieves the lowest RIS and DDI rates while maintaining superior accuracy. Since RIS is measured with respect to personalized risk boundaries, these results suggest the effectiveness of RAPO in better balancing therapeutic efficacy and safety under patient-specific safety constraints. 3) General reasoning baselines (e.g., Reason4Rec and EXP3RT) provide recommendation rationales but show limited overall performance. While these methods introduce interpretability, they fail to translate their reasoning capabilities into clinically accurate and safe decisions. This limitation may stem from the fact that these frameworks are adapted from general domains that prioritize user preference, and consequently lack explicit modeling of stringent safety constraints and heterogeneous clinical contexts.

Table 3: Ablation study of RES-MR on the MIMIC-III dataset.

Methods	Jaccard $\uparrow$	DDI $\downarrow$	BLEURT $\uparrow$
RES-MR w/o Reasoning	0.2414	0.1474	-
RES-MR w/o Retrieval	0.2244	0.1626	0.4117
RES-MR w/o RAPO	0.2668	0.1505	0.4408
RES-MR w/o Disentangle	0.2631	0.1211	0.4410
RES-MR w/o Dense Reward	0.2714	0.1103	0.4483
RES-MR w/o Sparse Reward	0.2631	0.1386	0.4396
RES-MR w/o Trade-off	0.2703	0.1324	0.4387
RES-MR w/o Clipping	0.2688	0.1086	0.4337
RES-MR-8B	0.2803	0.1031	0.4538

5.3 Ablation Study

To validate the effectiveness of each component in RES-MR, we design the following variants for comparative analysis: 1) w/o Reasoning: Fine-tunes on labels only, skipping the reasoning stage; 2) w/o Retrieval: Removes retrieved context from the input; 3) w/o RAPO: Removes the reinforcement learning stage; 4) w/o Disentangle: Uses raw LLM embeddings without dual-factor disentanglement; 5) w/o Dense/Sparse Reward: Removes dense (Efficacy/Risk) or sparse (Jaccard/DDI) rewards; 6) w/o Trade-off: Uses a fixed static sum instead of a dynamic trade-off; 7) w/o Clipping: Removes the reward clipping. The experimental results in Table 3 indicate several key findings. First, significant accuracy drops for w/o Retrieval and w/o Reasoning confirm that retrieval-augmented reasoning is fundamental for understanding complex clinical contexts and making accurate decisions. Second, the marked increase in DDI rates for w/o RAPO demonstrates that supervised learning alone is insufficient to internalize patient-specific risk boundary, necessitating RL for exploration and correction. Furthermore, the comprehensive performance decline in w/o Disentangle highlights that disentangled factors are critical for constructing high-quality prototypes. Finally, removing specific reward components or the Trade-off/Clipping mechanisms hinders the model from achieving

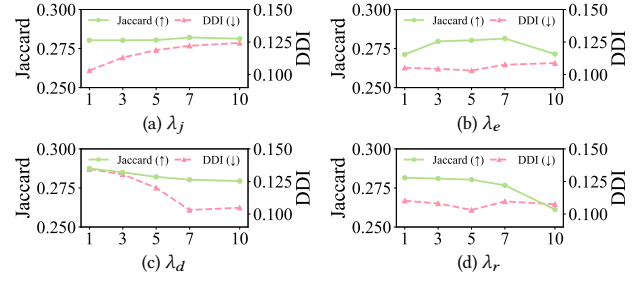


Figure 3: Hyperparameter effects on model performance.

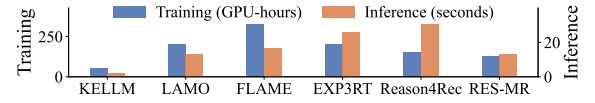


Figure 4: Efficiency analysis of training cost and inference time on MIMIC-III. Training uses 8× NVIDIA A100 GPUs, while inference uses a single NVIDIA A100 GPU.

an optimal balance between efficacy and safety (e.g., w/o Clipping may induce reward hacking, ignoring safety risks for high scores), highlighting the importance of our fine-grained reward design.

5.4 Hyperparameter Analysis

In this section, we analyze the impact of four key hyperparameters on the performance of RES-MR on the MIMIC-III dataset, i.e., Jaccard reward weight λ_j , efficacy bonus weight λ_e , DDI penalty weight λ_d , and risk penalty weight λ_r . We vary these parameters within the range of $\{1, 3, 5, 7, 10\}$ and report the results of Jaccard and DDI rate in Figure 3. As shown in Figure 3 (a) and (b), increasing efficacy-related weights generally encourages the model to better fit the target medication distribution. However, we observe that an excessively large λ_e leads to a slight decline in Jaccard. This suggests that over-emphasizing semantic similarity may divert the model towards “functionally similar” alternatives rather than the exact ground-truth labels, causing a drop in performance. Figure 3 (c) and (d) illustrate the trade-off between accuracy and safety. As λ_d and λ_r increase, DDI rates decrease significantly, validating the effectiveness of RAPO in suppressing unsafe behaviors. Notably, extremely high penalties cause a sharp collapse in Jaccard, reflecting an over-conservative policy that refuses prescriptions to avoid potential risks. Given that clinical safety is our primary consideration, we select weight settings that strictly control risks, even though this may incur a marginal loss in accuracy.

5.5 Efficiency Analysis

To assess practical deployability, we compare the computational costs of RES-MR with representative LLM-based baselines in Figure 4, with all experiments conducted in a unified hardware environment. While KELLM achieves the lowest time due to its non-generative discriminative design, RES-MR demonstrates superior efficiency among generative LLM-based methods. Specifically, regarding training, RES-MR requires approximately 125 GPU-hours, significantly lower than FLAME, which relies on a resource-intensive

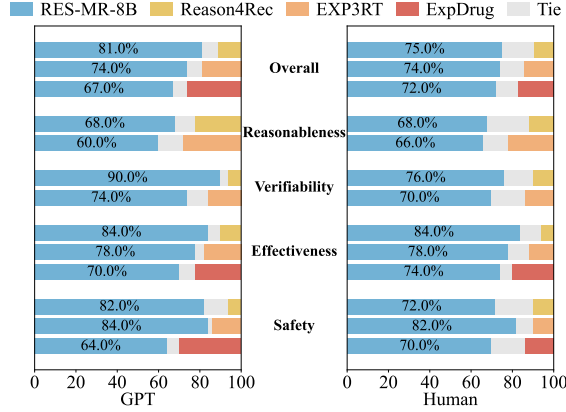


Figure 5: Human and GPT evaluations on pairwise quality comparison of explanations (p-value < 0.05).

Table 4: A case study of a patient in the MIMIC-III dataset.

Subject 41115 (80y, Male). Proc: AVR + CABG. Hx: AFib, HTN.	
Context	Verifiable Rationale → Recommendation
Surgical Protocol	Expl: “Given extracorporeal circulation , requires reversal of neuromuscular blockade...” → Rec: Neostigmine, Glycopyrronium
Stroke Risk	Expl: “Restart Warfarin post-op once hemostasis secured to prevent thromboembolism.” → Rec: Warfarin (<i>personalized risk boundary</i>)
Fluid Balance	Expl: “Replete K+ losses caused by aggressive diuresis to maintain electrical stability.” → Rec: Furosemide, Potassium Cl

multi-stage training pipeline. In terms of inference, unlike reasoning-enhanced baselines such as Reason4Rec and EXP3RT that incur high time overheads due to generating lengthy rationales, RES-MR achieves an inference time of only 13.3 s, comparable to the non-reasoning baseline LAMO. This efficiency is primarily attributed to the ECRD stage, which distills concise, clinically grounded reasoning paths, avoiding costly multi-pass generation during inference.

5.6 Explanation Quality Assessment

To evaluate explanation quality, we conduct a blind pairwise comparison on 50 randomly sampled MIMIC-III test instances, assessed by both GPT-5 and three human evaluators with medical or pharmaceutical backgrounds. The evaluation assesses the generated explanations via **Reasonableness** and **Verifiability**, and the recommended medication sets via **Effectiveness** and **Safety**. Note that ExpDrug is evaluated only on the latter two due to its lack of textual explanations. As shown in Figure 5, RES-MR consistently outperforms baselines. This superiority stems from ECRD, which enables the model to internalize verifiable, clinically grounded reasoning paths, and RAPO, which explicitly enforces safety via reward shaping. Moreover, human evaluators show substantial inter-rater agreement across all four dimensions (Fleiss’ $\kappa > 0.7$), and the agreement between human and GPT-5 judgments further supports the reliability of these findings.

5.7 Case Study

Table 4 exemplifies how RES-MR generates medication recommendations with **verifiable explanations**. Instead of relying on superficial co-occurrence, the model demonstrates deep reasoning: it adheres to procedural protocols by recommending Neostigmine and Glycopyrronium for neuromuscular blockade reversal, while leveraging pharmacological knowledge from *Knowledge Graph Retrieval* to anticipate Furosemide-induced hypokalemia, proactively co-prescribing Potassium Chloride. Furthermore, in addressing the conflict between stroke risk (AFib) and postoperative bleeding, RES-MR leverages fine-grained risk perception from RAPO to establish a **patient-specific risk boundary**. It prudently advises prescribing Warfarin only “once hemostasis is secured.” This nuance attests to the model’s ability to handle intricate clinical trade-offs while generating verifiable rationales and accurate, safe recommendations.

6 Conclusion

In this paper, we proposed RES-MR, a novel risk-aware reasoning framework designed to address the critical challenges of interpretability and personalized safety in medication recommendation. By integrating Explainable Clinical Reasoning Distillation (ECRD) for transparent reasoning and Risk-Aware Policy Optimization (RAPO) for dynamic safety balancing, RES-MR transforms recommendations into an explainable generative process. Experiments on three datasets demonstrate that RES-MR significantly outperforms state-of-the-art baselines, achieving a superior accuracy-safety trade-off. In future work, we will explore agentic reasoning workflows with iterative retrieval and self-verification to strengthen faithful clinical reasoning and further enhance decision reliability.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62376130), Program of New Twenty Policies for Universities of Jinan (No.202333008), and the Pilot Project for Integrated Innovation of Science, Education, Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01).

References

- [1] Zafar Ali, Yi Huang, Irfan Ullah, Junlan Feng, Chao Deng, Nimbeshaho Thierry, et al. 2023. Deep learning for medication recommendation: A systematic survey. *Data Intelligence* 5, 2 (2023), 303–354.
- [2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, et al. 2023. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.
- [3] Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, et al. 2021. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems* 40, 3 (2021), 1–23.
- [4] Suman Bhoi, Mong Li Lee, Wynne Hsu, and Ngai Chuan Tan. 2023. REFINE: A fine-grained medication recommendation system using deep learning and personalized drug interaction modeling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 24013–24024.
- [5] Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data* 10, 1 (2023), 1–16.
- [6] Lichang Chen, Chen Zhu, Jiu-hai Chen, Davit Soselia, Tianyi Zhou, et al. 2024. ODIN: Disentangled reward mitigates hacking in RLHF. In *Proceedings of the 41st International Conference on Machine Learning*. 7935–7952.
- [7] Qianyu Chen, Xin Li, Yujie Fang, and Mingzhong Wang. 2025. Advancing confidence calibration and quantification in medication recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 106–117.
- [8] Qianyu Chen, Xin Li, Kunan Geng, and Mingzhong Wang. 2023. Context-aware safe medication recommendations with molecular graph and DDI graph

- embedding. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 7053–7060.
- [9] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758* (2025), 1–19.
 - [10] Tianzhe Chu, Yuxiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*. 1–21.
 - [11] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, et al. 2023. Uncovering ChatGPT’s capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1126–1132.
 - [12] Zhuohang Dang, Minnan Luo, Jihong Wang, Chengyou Jia, Haochen Han, et al. 2025. Disentangled noisy correspondence learning. *IEEE Transactions on Image Processing* 34, 1 (2025), 2602–2615.
 - [13] Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the 12th International Conference on Learning Representations*. 1–14.
 - [14] Mingliang Dou, Jijun Tang, Prayag Tiwari, Yijie Ding, and Fei Guo. 2024. Drug–drug interaction relation extraction based on deep learning: A review. *Comput. Surveys* 56, 6 (2024), 1,33.
 - [15] Chenxiao Fan, Chongming Gao, Wentao Shi, Yaxin Gong, Zhao Zihao, et al. 2025. Fine-grained list-wise alignment for generative medication recommendation. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*. 1–17.
 - [16] Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, et al. 2025. Toward better EHR reasoning in LLMs: Reinforcement learning with expert attention guidance. *arXiv preprint arXiv:2508.13579* (2025), 1–18.
 - [17] Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, et al. 2025. Reason4Rec: Large language models for recommendation with deliberative user preference alignment. *arXiv preprint arXiv:2502.02061* (2025), 1–11.
 - [18] Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, et al. 2025. The aloe family recipe for open and specialized healthcare LLMs. *arXiv preprint arXiv:2505.04388* (2025), 1–82.
 - [19] Fan Gong, Meng Wang, Haofer Wang, Sen Wang, and Mengyue Liu. 2021. SMR: Medical knowledge graph embedding for safe medicine recommendation. *Big Data Research* 23, 1 (2021), 1–9.
 - [20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025), 1–11.
 - [21] Ming He, Yongjie Zheng, Changle Li, and Man Zhou. 2025. Improving the safety of medication recommendation via graph augmented patient similarity network. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 833–842.
 - [22] Health Care Financing Administration, National Center for Health Statistics, and Commission on Professional and Hospital Activities. 1989. *The international classification of diseases, 9th revision, clinical modification: ICD-9-CM*. Vol. 3. Commission on Professional and Hospital Activities.
 - [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*. 1–20.
 - [24] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. 2024. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5325–5344.
 - [25] Chengkai Huang, Shoujin Wang, Xianzhi Wang, and Lina Yao. 2023. Dual contrastive transformer for hierarchical preference modeling in sequential recommendation. In *Proceedings of the 46th international acm sigir conference on research and development in information retrieval*. 99–109.
 - [26] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alex Gayles, Aya Shammout, et al. 2023. MIMIC-IV, A freely accessible electronic health record dataset. *Scientific Data* 10, 1 (2023), 1.
 - [27] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, et al. 2016. MIMIC-III, A freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.
 - [28] Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems*. 1–30.
 - [29] Jieyong Kim, Hyunseo Kim, Hyunjin Cho, SeongKu Kang, Buru Chang, et al. 2025. Review-driven personalized preference reasoning with large language models for recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1697–1706.
 - [30] Taeri Kim, Jiho Heo, Hyunjoon Kim, and Sang-Wook Kim. 2025. HI-DR: Exploiting health status-aware attention and an EHR graph+ for effective medication recommendation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. 11950–11958.
 - [31] Taeri Kim, Jiho Heo, Hongil Kim, Kijung Shin, and Sang-Wook Kim. 2024. VITA: ‘Carefully chosen and weighted less’ is better in medication recommendation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. 8600–8607.
 - [32] Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*. 1–14.
 - [33] Erika Y Lee, Tara Gomes, Aaron M Drucker, Nick Daneman, Ayesha Asaf, et al. 2024. Oral antibiotics and risk of serious cutaneous adverse drug reactions. *Journal of the American Medical Association* 332, 9 (2024), 730–737.
 - [34] Jin Li, Huilin Gu, Shoujin Wang, Qi Zhang, Shui Yu, Chen Wang, Xiwei Xu, and Fang Chen. 2026. Towards Fair Large Language Model-based Recommender Systems without Costly Retraining. *arXiv preprint arXiv:2601.17492* (2026).
 - [35] Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 399–416.
 - [36] Yishuo Li, Zhufeng Shao, Weimin Chen, Shoujin Wang, Yuehan Du, and Wenpeng Lu. 2024. Significance-aware medication recommendation with medication representation learning. In *Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design*, Vol. 1. 1633–1638.
 - [37] Yishuo Li, Qi Zhang, Wenpeng Lu, Xueping Peng, Weiye Zhang, et al. 2025. Time-aware medication recommendation via intervention of dynamic treatment regimes. In *Proceedings of the 34th ACM on Web Conference*. 5161–5172.
 - [38] Yuhao Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, et al. 2025. G-Refer: Graph retrieval-augmented large language model for explainable recommendation. In *Proceedings of the 34th ACM on Web Conference*. 240–251.
 - [39] Jiahao Liu, Shengkang Gu, Dongsheng Li, Guangping Zhang, Mingzhe Han, et al. 2025. AgentCF++: Memory-enhanced LLM-based agents for popularity-aware cross-domain recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2566–2571.
 - [40] Jiahao Liu, Xueshuo Yan, Dongsheng Li, Guangping Zhang, Hansu Gu, Peng Zhang, et al. 2025. Improving LLM-powered recommendations with personalized information. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2560–2565.
 - [41] Qidong Liu, Zhaopeng Qiu, Xiangyu Zhao, Xian Wu, Zijian Zhang, et al. 2025. A contrastive pretrain model with prompt tuning for multi-center medication recommendation. *ACM Transactions on Information Systems* 43, 3 (2025), 1–29.
 - [42] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, et al. 2024. Large language model distilling medication recommendation Model. *arXiv preprint arXiv:2402.02803* (2024), 1–12.
 - [43] Wenpeng Lu, Kangjun Liu, Jianlei Wang, Xueping Peng, Tao Shen, Fa Zhu, Weiye Zhang, Jiabing Zhu, Tao Xin, and Athanasios V. Vasilakos. 2025. Advancing chinese conversation-based patient guidance with a benchmark and knowledge-evolvable assistant. *IEEE Journal of Biomedical and Health Informatics* 1, 1 (2025), 1–12.
 - [44] Wenpeng Lu, Yuteng Zhang, Shoujin Wang, Heyan Huang, Qian Liu, and Sheng Luo. 2020. Concept representation by learning explicit and implicit concept couplings. *IEEE Intelligent Systems* 36, 1 (2020), 6–15.
 - [45] Xuan Lu, Yanhong Hao, Furong Peng, Zheqing Zhu, and Zhanwen Cheng. 2025. ExpDrug: An explainable drug recommendation model based on space feature mapping. *Neurocomputing* 619 (2025), 1–11.
 - [46] Qiyao Ma, Xubin Ren, and Chao Huang. 2024. XRec: Large language models for explainable recommendation. In *Findings of the 29th Conference on Empirical Methods in Natural Language Processing*. 391–402.
 - [47] Costas Mavromatis and George Karypis. 2025. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. 16682–16699.
 - [48] Anton Nilsson, Jonas Björk, and Carl Bonander. 2022. Proxy variables and the generalizability of study results. *American Journal of Epidemiology* 192, 3 (2022), 448–454.
 - [49] Susana Nunes, Samy Badreddine, and Catia Pesquita. 2025. Rewarding explainability in drug repurposing with knowledge graphs. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*. 4624–4632.
 - [50] Yang Ouyang, Yuchen Wu, He Wang, Chenyang Zhang, Furui Cheng, et al. 2024. Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 1238–1248.
 - [51] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, et al. 2018. The eICU collaborative research database, A freely available multi-center database for critical care research. *Scientific Data* 5, 1 (2018), 180178.
 - [52] Sellam, Thibault and Das, Dipanjan and Parikh, Ankur. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7881–7892.
 - [53] Alyaa Hamel Sfayyih, Nasri Sulaiman, and Ahmad H Sabry. 2023. A review on lung disease recognition by acoustic signal analysis with deep learning networks. *Journal of Big Data* 10, 1 (2023), 101.
 - [54] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings*

- of the 28th International Joint Conference on Artificial Intelligence. 5953–5959.
- [55] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. GAMENet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 1126–1133.
- [56] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, et al. 2022. 4SDrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3970–3980.
- [57] Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, et al. 2024. Leveraging LLM reasoning enhances personalized recommender systems. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 13176–13188.
- [58] Boyuan Wang, Yun Qu, Yuhang Jiang, Jianzhun Shao, Chang Liu, et al. 2024. LLM-empowered state representation for reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*. 51348–51375.
- [59] Nan Wang, Shoujin Wang, Yan Wang, Quan Z Sheng, and Mehmet Orgun. 2020. Modelling local and global dependencies for next-item recommendations. In *International Conference on Web Information Systems Engineering*. Springer, 285–300.
- [60] Nan Wang, Shoujin Wang, Yan Wang, Quan Z Sheng, and Mehmet A Orgun. 2022. Exploiting intra-and inter-session dependencies for session-based recommendations. *World Wide Web* 25, 1 (2022), 425–443.
- [61] Shijie Wang, Wenqi Fan, Yue Feng, Lin Shanru, Xinyu Ma, et al. 2025. Knowledge graph retrieval-augmented generation for LLM-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. 27152–27168.
- [62] Shoujin Wang, Wentao Wang, Xiuzhen Zhang, Yan Wang, Huan Liu, and Fang Chen. 2024. A hierarchical and disentangling interest learning framework for unbiased and true news recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3200–3211.
- [63] Yanda Wang, Weitong Chen, Dechang Pi, Lin Yue, Sen Wang, et al. 2021. Self-supervised adversarial distribution regularization for medication recommendation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. 3134–3140.
- [64] Shoujin Wang, Guoqiang Zhang, Xu Zhang, Antai Ouyang, Xiaodong Zhang, Wenpeng Lu, Fanqing Meng. 2019. Graph-Based Chinese Word Sense Disambiguation with Multi-Knowledge Integration. *Computers, Materials & Continua* 61, 1 (2019), 197–212.
- [65] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, et al. 2018. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* 46, D1 (2018), D1074–D1082.
- [66] Feiyue Wu, Tianxing Wu, and Shenqi Jing. 2025. ARMR: Adaptively responsive network for medication recommendation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*. 7831–7839.
- [67] Jialun Wu, Yuxin Dong, Zeyu Gao, Tieliang Gong, and Chen Li. 2023. Dual attention and patient similarity network for drug recommendation. *Bioinformatics* 39, 1 (2023), 1–12.
- [68] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, et al. 2025. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826* (2025), 1–19.
- [69] Jialun Wu, Kai He, Rui Mao, Xuequn Shang, and Erik Cambria. 2025. Harnessing the potential of multimodal EHR Data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion* 123 (2025), 103283.
- [70] Jialun Wu, Buyue Qian, Yang Li, Zeyu Gao, Meizhi Ju, Yifan Yang, et al. 2022. Leveraging multiple types of domain knowledge for safe and effective drug recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2169–2178.
- [71] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the 31st ACM Web Conference*. 935–945.
- [72] Su Xian, Monika E Grabowska, Iftekhar J Kullo, Yuan Luo, Jordan W Smoller, et al. 2025. Transformer patient embedding using electronic health records enables patient stratification and progression analysis. *npj Digital Medicine* 8, 1 (2025), 521.
- [73] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 641–649.
- [74] Tianhan Xu and Bin Li. 2025. KELLM: Knowledge-enhanced label-wise large language model for safe and interpretable drug recommendation. *Electronics* 14, 1 (2025), 1–23.
- [75] Zhenhao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, et al. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2905–2909.
- [76] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change matters: Medication change prediction with recurrent residual networks. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. 3728–3734.
- [77] 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 30th International Joint Conference on Artificial Intelligence*. 3735–3741.
- [78] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. MoleRec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the 32nd ACM Web Conference*. 4075–4085.
- [79] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476* (2025), 1–16.
- [80] Haodi Zhang, Jiawei Wen, Jiahong Li, Yuanfeng Song, Liang-Jie Zhang, et al. 2025. SSPNet: Leveraging robust medication recommendation with history and knowledge. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*. 9465–9473.
- [81] Qi Zhang, Yayi Yang, Chongyang Shi, An Lao, Liang Hu, Shoujin Wang, and Usman Naseem. 2024. Rumor Detection With Hierarchical Representation on Bipartite Ad Hoc Event Trees. *IEEE Transactions on Neural Networks and Learning Systems* 35, 10 (2024), 14112–14124.
- [82] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1315–1324.
- [83] Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2023. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems* 41, 1 (2023), 1–24.
- [84] Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li. 2024. Enhancing precision drug recommendations via in-depth exploration of motif relationships. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 8164–8178.
- [85] Zihao Zhao, Chenxiao Fan, Chongming Gao, Fuli Feng, and Xiangnan He. 2025. Fine-grained alignment of large language models for general medication recommendation without overprescription. *arXiv preprint arXiv:2503.03687* (2025), 1–17.
- [86] Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, et al. 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*. 533–542.
- [87] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, et al. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071* (2025), 1–7.
- [88] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, et al. 2023. Interaction-aware drug package recommendation via policy gradient. *ACM Transactions on Information Systems* 41, 1 (2023), 1–32.
- [89] Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, Jae Boum Kim, Shoujin Wang, Raymond Chi-Wing Wong, and Sunghun Kim. 2025. When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study. In *Proceedings of the ACM on Web Conference 2025*. 275–292.