Awaken the Giant: Activating LLMs via Deep Model Guidance for Boundary-aware Medication Recommendation

Hang Lv
Fuzhou University
College of Computer and Data Science, Fuzhou, China
lvhangkenn@gmail.com

Zixuan Guo
Fuzhou University
Maynooth International Engineering College, Fuzhou, China
832304221@fzu.edu.cn

Yanchao Tan*
Fuzhou University
College of Computer and Data Science, Fuzhou, China
yctan@fzu.edu.cn

Wanzi Shao
Fuzhou University
College of Computer and Data Science, Fuzhou, China
shaowz0302@163.com

Hengyu Zhang
Macquarie University
School of Computing
Sydney, Australia
hengyu.zhang3@hdr.mq.edu.au

Carl Yang
Emory University
Department of Computer Science
Atlanta, USA
j.carlyang@emory.edu

Abstract

Accurate and safe medication recommendations from Electronic Health Records (EHRs) are essential for clinical decision support. While Large Language Models (LLMs) have shown strong semantic reasoning capabilities in healthcare, they tend to make coarse binary predictions, overlooking medications near the decision boundary and leading to overprescription. In contrast, deep models offer fine-grained probability outputs but lack contextual reasoning needed for complex boundary cases. To address this, we propose a boundary-aware medication recommendation framework (`GiantMed`) that activates the potential of LLM "giant" under deep model guidance. Specifically, `GiantMed` leverages a deep model to identify boundary medications and directs the LLM to focus on these clinically ambiguous yet informative cases. Furthermore, we augment contextual knowledge of boundary medications by retrieving relevant historical EHRs and integrating Drug-Drug Interaction (DDI) constraints. The final recommendation is generated by incorporating the LLM-refined boundary medications with confident predictions from the deep model. Extensive experiments on two real-world EHR datasets demonstrate that `GiantMed`[1] achieves state-of-the-art accuracy while reducing DDI rates.

CCS Concepts

• **Applied computing → Health informatics**; • **Information systems → Data mining**; • **Computing methodologies → Artificial intelligence**.

Keywords

Medication Recommendation, Large Language Models, Boundary-aware Medication

ACM Reference Format:
Hang Lv, Zixuan Guo, Yanchao Tan, Wanzi Shao, Hengyu Zhang, and Carl Yang. 2026. Awaken the Giant: Activating LLMs via Deep Model Guidance for Boundary-aware Medication Recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770854.3780297

1 Introduction

Medication recommendation is essential for clinical decision-making, aiming to provide accurate and safe prescriptions tailored to a patient's condition based on Electronic Health Records (EHRs) [38, 39, 42]. With the rapid advancement of Large Language Models (LLMs) such as GPT [24] and LLaMA [7] series, recent research has started exploring their potential in healthcare (e.g., clinical report generation [13] and medical question answering [14]), leveraging their strong semantic reasoning and generalization abilities [12, 43]. These strengths have motivated adaptations of LLMs to medication recommendation [9, 21], typically by fine-tuning them on large-scale clinical corpora to enhance domain alignment.

Despite their potential, LLM-based approaches still face critical challenges in medication recommendation [9, 43]. As illustrated in Figure 1(a)-top, LLMs like Qwen3-8B [36] often make coarse binary predictions about whether a medication should be recommended or not, thereby overlooking those near the decision boundary. This behavior frequently leads to overprescribing and reduces clinical precision, as illustrated in Figure 1(b), where LLMs recommend more medications than the ground-truth set. In contrast, deep models trained on structured EHR data (e.g., SafeDrug [38]) offer fine-grained probability distributions over candidate medications (Figure 1(a)-bottom). This enables more precise control over predictions and improves coverage of medications within the boundary region, which tends to be missed by LLMs. Such boundary medications are often ambiguous yet clinically informative, and identifying them correctly is critical for generating safe and effective prescriptions. Nevertheless, deep models often lack the contextual reasoning required to handle complex cases, especially when boundary medications are involved.

While LLMs offer semantic reasoning capabilities but suffer from coarse binary decision behavior, deep models provide fine-grained

*Yanchao Tan is the corresponding author.
[1] https://github.com/lvhangkenn/GiantMed

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3780297

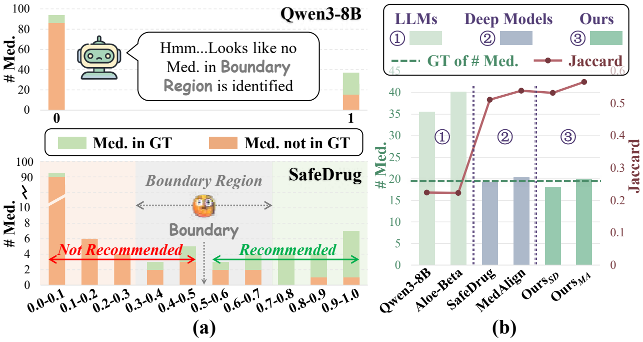


Figure 1: (a) Medication probability distributions for a patient's visit from Qwen3-8B (LLM) and SafeDrug (deep model). The LLM fails to capture medications within the boundary region, while the deep model offers fine-grained probabilities around the boundary. (b) Comparison of different methods on MIMIC-III by the average number of recommended medications (bars) and Jaccard similarity (line). Method types are color-coded. "Med." is medications and "GT" is ground truth. "Jaccard" quantifies the similarity between the predicted and ground-truth medication sets.

control yet lack the reasoning needed for complex clinical scenarios. This complementary pattern of limitations raises a key question: *Can we integrate the accurate control from deep models with the semantic reasoning ability of LLMs to improve medication recommendation, especially for boundary cases?*

To this end, we propose `GiantMed`, a boundary-aware medication recommendation framework that activates the potential of the LLM "giant" through deep model guidance. Rather than relying on LLMs to make broad and unguided predictions, `GiantMed` strategically directs its attention toward ambiguous yet clinically important medications near the decision boundary. As shown in Figure 1(b), `GiantMed` variants (e.g., Ours$_{SD}$ and Ours$_{MA}$, respectively guided by SafeDrug [38] and MedAlign [23]), consistently improve the recommendation accuracy while alleviating overprescribing issues.

Specifically, we first leverage a trained deep model's fine-grained probability outputs to divide all medications into boundary and confident subsets for a patient's visit, extracting those within the boundary region for further refinement by LLMs. We then augment contextual knowledge of boundary medications by retrieving relevant historical EHRs and applying Drug-Drug Interaction (DDI) constraints. Finally, we incorporate the refined boundary medication probabilities from the LLM with the confident predictions from the deep model to generate more accurate and safer prescriptions.

The main contributions of our work are summarized as follows:

- *Formulation of Boundary-aware Medication Recommendation.* We introduce a novel formulation that activates the reasoning capabilities of LLM "giant" on medications within the boundary region. By focusing on ambiguous yet clinically informative cases, our approach mitigates coarse binary decision behaviors and enables more fine-grained, precise, and safer recommendations.
- *Deep Model-Guided LLM Framework.* We propose `GiantMed`, a hybrid framework that guides LLMs using deep model predictions to identify boundary medications. These are further refined via a boundary-aware prompt, augmented with retrieved EHRs and integrated DDI constraints, effectively integrating the strengths of deep models and LLMs for medication recommendation. Notably, `GiantMed` is a model-agnostic plug-in framework that can flexibly adapt to different guiding deep models. This design effectively integrates probabilistic outputs from state-of-the-art deep models to improve accuracy and reliability by targeting medications near the decision boundary.
- *Extensive Experimental Validation.* We conduct comprehensive experiments on two real-world EHR datasets, demonstrating that `GiantMed` significantly outperforms state-of-the-art baselines in both accuracy and safety (minimizing DDIs).

2 Related Work

2.1 Medication Recommendation

Medication recommendation aims to provide accurate and safe prescriptions for patients via personalized treatment [27, 29, 42]. Existing medication recommendation methods can be broadly categorized into instance-based and longitudinal approaches, based on how patient information is modeled. Instance-based methods [5, 25] typically rely on structured features extracted from a single patient visit. However, such methods often overlook valuable historical patient data. In contrast, longitudinal methods [20, 26, 35, 37] model the long-term progression of diseases by integrating temporal information from patients' hospitalization histories. For instance, SafeDrug [38] introduced dual molecular graph encoders to leverage drug structural information, thereby improving the accuracy and safety of recommendations. RETAIN [6] employed a dual-layer recurrent neural network with a reverse-time attention mechanism to identify previous visits and key clinical features that have a significant impact on the current condition. VITA [19] improved patient state modeling by selectively filtering irrelevant historical visits and employing target-aware attention to precisely quantify their relevance to the current clinical context. Moreover, MedAlign [23] further explored the correlations and complementary information among different medication modalities to optimize recommendation performance. While existing deep models effectively leverage structured clinical data for enhanced recommendations, they lack the contextual reasoning capabilities, particularly for complex scenarios, thereby failing to handle ambiguous medications.

To address these limitations, recent advances in Large Language Models (LLMs) have opened up new opportunities for medication recommendation. For example, LAMO [43] fine-tuned LLMs using structured data and unstructured clinical notes, enabling it to better capture patient conditions through enhanced semantic understanding. FLAME [9] generated prescription drug-by-drug via group relative policy optimization over LLMs, integrating multi-source medical knowledge through hybrid patient representations. Despite their promising performance, these LLM-based methods often require fine-tuning on large-scale clinical corpora, leading to high computational costs and slow inference.

2.2 LLM-augmented Healthcare

Recent advancements in artificial intelligence, particularly LLMs, have demonstrated impressive performance on natural language processing tasks [4, 32]. CoAD [31] highlighted LLMs' immense potential in healthcare, where they are increasingly applied to tasks such as medical question answering [14], clinical report generation [13], and medication recommendation [28]. The sophisticated understanding, generation, and reasoning capabilities of LLMs can help enhance diagnostics and offer more tailored treatment options. General-purpose models like GPT [24] and LLaMA [7] series have shown superior performance across various healthcare-related tasks [12], motivating the development of specialized medical LLMs such as BioGPT [22] and Aloe [10].

To tackle specific clinical prediction challenges, researchers have further fine-tuned LLMs on large-scale medical data, aiming to improve performance in specialized downstream applications. For example, CPLLM [3] leveraged historical diagnosis records to predict future disease diagnoses and hospital readmissions. Health-LLM [40] analyzed patient health reports to deliver personalized advice by combining large-scale feature extraction with medical knowledge. Similarly, LLM-DG [17] used an LLM-enhanced mechanism to semantically enrich patient, disease, and discharge summary data, thereby improving prediction accuracy by integrating medical knowledge and capturing similar disease trajectories. However, applying LLMs to medication recommendation presents unique challenges due to their coarse binary decision behavior, which can even lead to the risk of overprescribing [9, 43]. Consequently, how to effectively harness LLMs to generate accurate and safe medication prescriptions with fine-grained control remains largely unexplored.

3 Theoretical Motivation: Boundary-aware Medication Recommendation

In medication recommendation tasks, accurately identifying boundary medications (i.e., those within the boundary region near the decision boundary) is crucial for enhancing both the precision and safety of prescriptions. As shown in Figure 1(a), deep models naturally predict fine-grained medication probability distributions, which facilitates the definition of a boundary-based prediction space to distinguish between boundary and confident decisions. In contrast, Large Language Models (LLMs) often generate coarse binary predictions on whether a medication should be recommended or not, which can lead to neglecting medications near the decision boundary and overprescribing.

To formalize this observation, we ground our proposed approach in margin-based generalization theory [1, 2, 30]. Let $\mathcal{V}$ denote the input space (e.g., a patient's visit), and $\mathcal{Y} = \{0, 1\}$ be the label space indicating whether a medication should be recommended. Given a real-valued scoring function $f : \mathcal{V} \to \mathbb{R}$ that predicts the likelihood of medication recommendation (e.g., SafeDrug [38]), the predicted label for a specific input $v \in \mathcal{V}$ is:

$$\hat{y} = \mathbb{I}[f(v) > \delta], \tag{1}$$

where $f(v) \in \mathbb{R}$ is the predicted score, δ is the decision boundary, and $\mathbb{I}[\cdot]$ denotes the indicator function that returns 1 if the medication is recommended, and 0 otherwise. The prediction margin for an example (v, y) is defined as follows:

$$\gamma(v, y) = (2y - 1) \cdot f(v), \tag{2}$$

where $y \in \mathcal{Y}$ is the ground-truth label. A larger margin $\gamma(v, y)$ implies higher prediction confidence, while smaller margins correspond to boundary cases.

To analyze the generalization behavior of such scoring functions, we consider the broader real-valued function class $\mathcal{F}$ with bounded norm (e.g., linear classifiers or regularized neural networks). The following generalization bound [2, 30] holds for any $f \in \mathcal{F}$:

Lemma 1. ***(Margin-based Generalization Bound)****. Let $\gamma > 0$ be a margin boundary. With high probability over a sample of size n, the generalization error $R(f)$ satisfies:*

$$R(f) \leq \widehat{R}\gamma(f) + O\left(\frac{1}{\gamma} \cdot \sqrt{\frac{C(\mathcal{F})}{n}}\right), \tag{3}$$

where $\widehat{R}\gamma(f)$ is the empirical fraction of margin violations (i.e., samples with $\gamma(v, y) \leq \gamma$), and $C(\mathcal{F})$ is the function class complexity.

This bound highlights the importance of minimizing the number of low-margin samples to improve generalization. The detailed proof can be found in [2, 30]. For medication recommendation, $\widehat{R}_\gamma(f)$ corresponds to the proportion of boundary medications (i.e., those with predicted scores falling within a margin region around the decision boundary δ). Therefore, effectively identifying and refining such cases is critical for improving the accuracy and safety of medication recommendations. Importantly, the above analysis is agnostic to the specific model architecture and only requires access to real-valued prediction scores.

However, LLMs output coarse binary predictions without probability margins, limiting their ability to recognize clinically ambiguous yet important cases near the boundary δ. This may lead to overprescription and impair safety. Motivated by this, our GiantMed leverages deep models' fine-grained predictions to identify boundary medications and activate the potential of LLM "giant". This design focuses the semantic reasoning power of LLMs on boundary cases, thereby enhancing generalization and efficiency without requiring full-scale inference over all medications.

4 Methodology

4.1 Problem Formulation and Overview

Our proposed GiantMed aims to provide accurate and safe medication recommendations based on Electronic Health Records (EHRs) and Drug-Drug Interaction (DDI) matrix. Each patient's EHR consists of a sequence of visits, where each visit contains a set of diagnoses, procedures, and prescribed medications. Since our GiantMed directly performs inference based on the patient's visit using a Large Language Model (LLM) and a trained deep medication recommendation model, we omit the temporal index and represent the current visit as v in the following sections for simplicity.

Figure 2 provides an overview of the three core components of our proposed GiantMed.

- *Boundary Medication Set Identification via Deep Model Guidance:* We leverage a trained deep model (e.g., SafeDrug [38]) to predict fine-grained medication probabilities for the patient's visit, partitioning all medications into boundary and confident subsets.
- *Boundary Medication Augmentation with Retrieved EHRs and DDI Constraints:* We retrieve clinically relevant historical EHRs most similar to the patient's visit and extract DDI constraints over the boundary medication set, thereby augmenting contextual knowledge to support subsequent refinement by LLMs.
- *Boundary-aware Medication Recommendation via LLMs:* We construct a synthesized boundary-aware prompt to activate the

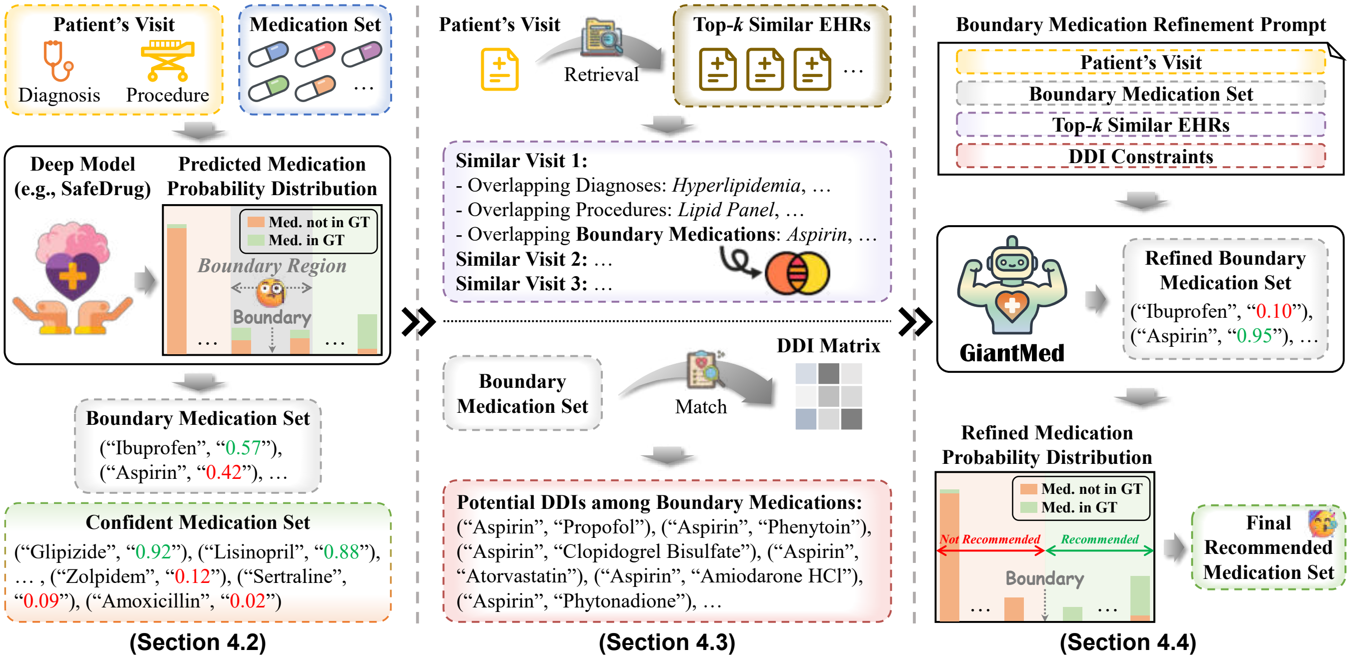


Figure 2: The overall framework of our GiantMed, which consists of three core modules: Boundary Medication Set Identification via Deep Model Guidance (Section 4.2), Boundary Medication Augmentation with Retrieved EHRs and DDI Constraints (Section 4.3), and Boundary-aware Medication Recommendation via LLMs (Section 4.4).

semantic reasoning potential of the LLM (e.g., Qwen3-8B [36]), enabling it to refine boundary medication probabilities and generate the final recommendation.

4.2 Boundary Medication Set Identification via Deep Model Guidance

While LLMs have demonstrated strong reasoning capabilities in healthcare, their application to medication recommendation remains limited by coarse binary decision behavior. Faced with a large set of candidate medications, LLMs tend to ignore medications near the decision boundary, leading to overprescription [9, 43]. Such overgeneralized prescriptions not only reduce recommendation accuracy but also lead to potential clinical risks.

To mitigate this issue, we introduce a deep model-guided mechanism that directs the LLM's semantic reasoning toward the ambiguous yet informative boundary medications. Specifically, we utilize a trained deep medication recommendation model (e.g., SafeDrug [38]) to predict the probability distributions over candidate medications for each patient's visit. Based on these fine-grained predictions, we divide all medications into boundary and confident subsets. The extracted boundary medication set is further refined by LLMs in Section 4.4.

4.2.1 Medication Probability Prediction. A deep medication recommendation model backbone can be viewed as a trained multi-label classifier f. Given the patient's visit $\boldsymbol{v} = [\boldsymbol{d}, \boldsymbol{p}]$, the medication probability distribution $\boldsymbol{y}$ is predicted as follows:

$$\boldsymbol{y} = f(\boldsymbol{v}) = f([\boldsymbol{d}, \boldsymbol{p}]) = [y_1, y_2, \ldots, y_{|\mathcal{M}|}], \tag{4}$$

where $\boldsymbol{d} \in \{0,1\}^{|\mathcal{D}|}$ and $\boldsymbol{p} \in \{0,1\}^{|\mathcal{P}|}$ are multi-hot vectors of the visit's diagnoses and procedures. $|\mathcal{D}|$, $|\mathcal{P}|$, and $|\mathcal{M}|$ denote the number of all candidate diagnoses, procedures, and medications. Particularly, medications with a predicted probability $y_i \geq \delta$ are directly recommended as part of the medication prescription, where δ is a decision boundary typically set to 0.5 [38, 39, 42].

4.2.2 Boundary and Confident Medication Set Extraction. Despite their powerful semantic understanding capabilities, LLMs often suffer from overprescription due to coarse decision control and a broad candidate medication space, limiting both recommendation accuracy and the effective use of reasoning potential. Inspired by the ability of deep models to capture medications around the boundary region based on predicted fine-grained probabilities (shown in Figure 1(a)), we employ deep model guidance to identify a medication subset that is most worthy of refinement by LLMs.

To be specific, we first define a boundary region centered around the recommendation boundary δ as $[\delta - \epsilon, \delta + \epsilon]$, where ϵ represents the width of the boundary region (we further investigate the impact of ϵ in Section 5.4). Subsequently, we divide all candidate medications $\mathcal{M}$ into two disjoint subsets based on their predicted probabilities from the deep model:

- Boundary Medication Set $\mathcal{M}_{bound} = \{m_i \mid y_i \in [\delta - \epsilon, \delta + \epsilon]\}$ with the corresponding predicted probability set $\mathcal{Y}_{bound}$.
- Confident Medication Set $\mathcal{M}_{confid} = \{m_j \mid y_j \in [0, \delta - \epsilon) \cup (\delta + \epsilon, 1]\}$ with the corresponding predicted probability set $\mathcal{Y}_{confid}$.

Notably, a medication m_i belongs to the boundary medication set $\mathcal{M}_{bound}$ when its predicted probability y_i falls within the boundary region. These medications, which deep models struggle to handle due to limited contextual understanding, require further semantic reasoning, making them ideal candidates for refinement by LLMs.

Finally, we textualize each medication in $\mathcal{M}_{bound}$ with its predicted probability via the prompt (described in the bottom left of Figure 2).

4.3 Boundary Medication Augmentation with Retrieved EHRs and DDI Constraints

As mentioned in Section 1, boundary medications are inherently ambiguous, as they often lack sufficient contextual support when relying solely on the information from a single patient visit in the EHR. Without access to broader clinical context or external medical knowledge, LLMs struggle to accurately interpret and reason about these boundary cases, reducing the recommendation accuracy.

To address this, we design a dual-perspective knowledge augmentation strategy that enriches contextual information and constrains unsafe predictions. Specifically, we retrieve clinically relevant EHRs to supplement contextual information for the patient's visit. Moreover, we integrate DDI knowledge over the extracted boundary medication set (cf., Section 4.2.2) to constrain unsafe prescriptions.

4.3.1 Similar EHR Retrieval for Clinical Augmentation. To provide relevant clinical knowledge, we retrieve the top-k EHRs most similar to the patient's visit. This enriches the contextual information available to LLMs, offering more focused guidance for further refining boundary medication recommendations [18, 34]. We compute the Jaccard similarity based on diagnosis overlap to quantify the relevance between the patient's visit v and each historical visit h in the training set as follows:

$$s_{diag}(v,h) = \frac{|\mathcal{D}_v \cap \mathcal{D}_h|}{|\mathcal{D}_v \cup \mathcal{D}_h|}, \tag{5}$$

where $\mathcal{D}_v$ and $\mathcal{D}_h$ are the diagnosis sets. We then rank all historical visits by $s_{diag}(v,h)$ and select the top-k most similar visits as contextual knowledge augmentation. Additional analyses using other clinical features (e.g., procedures) are detailed in Section 5.5. Furthermore, we textualize the retrieved similar visit EHRs using the simplified prompt template (illustrated in the top middle of Figure 2). Notably, we only include information overlapping with the patient's visit (i.e., shared diagnoses, procedures, and boundary medications) to highlight relevant contextual knowledge while keeping the prompt concise.

4.3.2 DDI Constraint Construction over Boundary Medications. To ensure the safety of recommended medications, we further construct DDI constraints for LLMs based on a predefined symmetric binary DDI adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{M}|}$, where $\boldsymbol{A}_{ij} = 1$ indicates an interaction between medications i and j. Given the extracted boundary medication set $\mathcal{M}_{bound}$ and DDI adjacency matrix $\boldsymbol{A}$, we match all potential DDI conflict set $C_{ddi} = \{(m_i, m_j) \mid \{m_i, m_j\} \subseteq \mathcal{M}_{bound}, \boldsymbol{A}_{ij} = 1\}$. The obtained DDI constraints are subsequently converted into textual format using the prompt (described in the bottom middle of Figure 2).

4.4 Boundary-aware Medication Recommendation via LLMs

With the boundary medication set (cf., Section 4.2.2) and the knowledge augmentation from retrieved EHRs and DDI constraints (cf., Section 4.3), we integrate the patient's visit information to complete the boundary-aware prompt construction for LLMs. Subsequently, the LLM (e.g., Qwen3-8B [36]) refines the probabilities of boundary medications and combines them with the confident predictions from the deep model to generate the final recommendation.

Our GiantMed framework activates the latent reasoning capabilities of LLMs with fine-grained control, directing their attention to medications within the boundary region. This refinement effectively enhances the decision behavior of LLMs, improving recommendation accuracy and safety while alleviating overprescribing. Notably, this design can flexibly and efficiently incorporate different LLMs with various deep model guidance without any task-specific fine-tuning. We further demonstrate the adaptability and efficiency of our GiantMed through comprehensive experiments in Section 5.2, Section 5.3, and Section 5.6.

4.4.1 Boundary Medication Probability Refinement. We synthesize the boundary medication refinement prompt $\mathcal{T}_v$ tailored to the patient's visit v via the following simplified template:

> Prompt for Boundary Medication Refinement
>
> **Patient's Visit:** [*Including diagnoses and procedures*]
> **Boundary Medication Set:** [*Including medications with probability predicted by the deep model*]
> **Top-k Similar EHRs:** [*Including overlapping diagnoses, procedures, and boundary medications*]
> **DDI Constraints:** [*Including potential side effects among boundary medications*]
>
> **Instruction:** Based on the above information, refine the predicted probabilities of boundary medications.
> **Output Format:** `(“Medication 1”, “Probability 1”), (“Medication 2”, “Probability 2”), ...`

Then, the LLM processes the prompt and outputs the refined boundary medication probabilities as follows:

$$\widehat{\mathcal{Y}}_{bound} = \mathrm{LLM}(\mathcal{T}_v). \tag{6}$$

The output is re-evaluated by the LLM based on the patient's health conditions and relevant clinical knowledge, ensuring focused and interpretable refinement.

4.4.2 LLM-based Medication Recommendation. We integrate the refined boundary medication probabilities $\widehat{\mathcal{Y}}_{bound}$ from the LLM with the confident medication probabilities $\mathcal{Y}_{confid}$ from the deep model (cf., Section 4.2.2), thereby generating the final recommended medication set as follows:

$$\widehat{\mathcal{M}} = \{m_i \mid y_i \in [\delta, 1]\}, \tag{7}$$

where δ is a recommendation boundary. By utilizing deep model guidance to activate LLMs, our GiantMed effectively and efficiently harnesses their strong semantic reasoning abilities, enabling accurate and safe boundary-aware medication recommendations.

4.5 Complexity Analysis

Our proposed GiantMed is a lightweight and efficient framework, which avoids fine-tuning LLMs for medication recommendation and instead relies on: (1) A single-pass inference from the trained deep model f, with time complexity $O(|\mathcal{D}| + |\mathcal{P}|)$, where $|\mathcal{D}|$ and $|\mathcal{P}|$ are the total numbers of all candidate diagnoses and procedures,

Table 1: Statistics of the datasets used in our experiments.

Dataset	MIMIC-III	MIMIC-IV
# of visits / # of patients	15,031 / 6,350	163,877 / 61,264
# of Diag. / # of Proc.	1,903 / 1,409	2,000 / 11,056
# of Med.	131	131
Avg./Max # of visits	2.3671 / 29	2.6749 / 70
Avg./Max # of Diag. per visit	10.2266 / 127	8.2343 / 270
Avg./Max # of Proc. per visit	3.8244 / 50	2.3579 / 95
Avg./Max # of Med. per visit	11.4361 / 65	6.5055 / 72
DDI Rate	0.0815	0.0793

respectively; (2) Relevant EHR retrieval via Jaccard similarity over diagnosis sets, with time complexity $O(k \cdot |\mathcal{D}|)$ for selecting top-k similar historical visits; (3) DDI constraint extraction over the boundary medication set $\mathcal{M}_{bound}$ using a binary DDI adjacency matrix $\boldsymbol{A}$, with time complexity $O(|\mathcal{M}_{bound}|^2)$ due to pairwise matching; (4) LLM inference with time complexity proportional to the prompt token length $O(|\mathcal{T}_v|)$. We construct a concise and focused prompt containing only the boundary medication set and relevant clinical knowledge, significantly reducing input token length and response time.

5 Experiments

In this section, we evaluate our GiantMed framework, focusing on the following six key research questions:

- **RQ1:** How does the proposed GiantMed compare with state-of-the-art methods in medication recommendation?
- **RQ2:** How does the choice of different backbone LLMs affect the performance of GiantMed?
- **RQ3:** How do key hyperparameters impact recommendation performance, and how to select their optimal values?
- **RQ4:** How do different boundary medication augmentation strategies affect the performance of GiantMed?
- **RQ5:** How does GiantMed perform in terms of efficiency?
- **RQ6:** How does GiantMed activate LLMs to enhance boundary-aware medication recommendation and generate explanations?

5.1 Experimental Settings

5.1.1 Datasets and Evaluation Metrics. We use two real-world Electronic Health Record (EHR) datasets to verify the effectiveness of GiantMed, i.e., **MIMIC-III** [15] and **MIMIC-IV** [16]. Both datasets are fully anonymized and carefully sanitized before our access. Following [23, 38, 39, 42], we chose patients who made at least two visits for both datasets and the ATC third-level code as the target label. The statistics are summarized in Table 1. Details on datasets are provided in Appendix A.

For evaluation metrics, we use Jaccard Similarity Score (Jaccard), Average F1 Score (F1), Precision Recall AUC (PRAUC), Drug-Drug Interaction Rate (DDI), and Average Number of Medications (# Med.), indicating how well the model aligns with real-world prescribing patterns, which are consistent with [26, 39, 42].

5.1.2 Methods for Comparison. To comprehensively evaluate the effectiveness of GiantMed, we adopt 14 representative state-of-the-art methods as baselines across three categories:

- **LLM-based methods:** We compare both general-purpose and domain-adapted LLMs, including **Qwen3-8B** [36], **LLaMA3.1-8B-Instruct** [7] (abbr. LLaMA3.1), and **LLaMA3.1-Aloe-Beta-8B** [10] (abbr. Aloe-Beta). Notably, we exclude LAMO [43] and FLAME [9] in our comparison, due to the unavailability of pre-trained checkpoints and training data;
- **Instance-based methods:** Consistent with prior works [38, 39], we select **LR** [5] and **ECC** [25] for comparison;
- **Longitudinal methods:** Follow [23, 38, 42], we include **RETAIN** [6], **LEAP** [41], **GAMENet** [26], **MICRON** [37], **VITA** [19], **SafeDrug** [38], **MoleRec** [39], **DEPOT** [42], and **MedAlign** [23].

5.1.3 Implementation Details. We split training, validation, and test sets by 2/3, 1/6, and 1/6, consistent with [23, 38, 39, 42]. For our GiantMed, we adopt the Qwen3-8B model as the default LLM backbone. The maximum input length for the LLM is set to 4096, and the temperature is fixed at 0.7. We set the boundary δ as 0.5 and the boundary region width ϵ as 0.2. For each patient's visit, we retrieve the top-k most similar EHRs (with $k = 3$) to enrich the contextual information in the prompt. We evaluate each model with the 5-fold cross-validation strategy. Both the mean and standard deviation of test performance are reported. All experiments are conducted on two NVIDIA RTX 3090 Ti GPUs. The full code for this work is available[2].

For LLM-based methods, we load their pretrained checkpoints from HuggingFace[3][4][5]. Notably, PRAUC is marked as "–" for LLM-based methods because they output binary decisions without probability scores, making precision-recall evaluation unavailable. This further highlights the benefit of GiantMed in enabling probability-aware refinement.

For instance-based and longitudinal methods, we optimize the compared baselines with the standard Adam optimizer and carefully tune their hyperparameters as suggested in the original papers. We set the embedding dimension to 64 and the batch size to 32.

5.2 Overall Performance Comparison (RQ1)

In this subsection, we present a comprehensive performance analysis of our proposed GiantMed framework compared to 14 representative baselines. Overall, our GiantMed demonstrates superior performance, attributed to its novel framework that integrates a deep model's fine-grained predictive guidance with an LLM's advanced semantic reasoning. As shown in Table 2, GiantMed consistently outperforms all baselines, achieving an average improvement of 3.74% in accuracy metrics and 4.10% reduction in DDI rate. By focusing on the refinement of clinically ambiguous yet informative boundary medications, GiantMed effectively balances between predictive accuracy and safety.

Compared with LLM-based methods. LLM-based approaches (e.g., Qwen3-8B and Aloe-Beta) often suffer from coarse binary decision behaviors, resulting in low accuracy and frequent overprescription. For instance, Qwen3-8B recommends over 35 medications per patient on average in MIMIC-III, almost twice the ground-truth

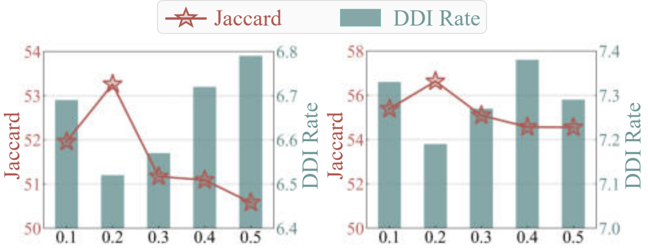
[2] https://github.com/lvhangkenn/GiantMed
[3] https://huggingface.co/Qwen/Qwen3-8B
[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5] https://huggingface.co/HPAI-BSC/Llama3.1-Aloe-Beta-8B

Table 2: Experimental results (%) on two EHR datasets. The best performances are highlighted in boldface, indicating statistically significant improvements according to the Wilcoxon signed-rank test, while the second-best results are underlined. Ground-truth # of Med. on the test sets is 19.79 for MIMIC-III and 11.98 for MIMIC-IV, respectively. GiantMed$_{SD}$, GiantMed$_{MR}$, GiantMed$_{DP}$, and GiantMed$_{MA}$ guided by SafeDrug, MoleRec, DEPOT, and MedAlign, respectively.

Dataset	MIMIC-III					MIMIC-IV				
Method	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓	# Med.	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓	# Med.
Qwen3-8B	$22.43_{\pm 2.05}$	$35.83_{\pm 2.71}$	–	$8.67_{\pm 0.24}$	35.53	$17.66_{\pm 1.72}$	$28.95_{\pm 2.83}$	–	$8.49_{\pm 0.18}$	23.81
LLaMA3.1	$19.71_{\pm 1.96}$	$31.70_{\pm 1.37}$	–	$8.83_{\pm 0.19}$	46.55	$14.13_{\pm 1.38}$	$23.39_{\pm 2.09}$	–	$7.91_{\pm 0.21}$	48.75
Aloe-Beta	$22.33_{\pm 1.34}$	$35.65_{\pm 2.13}$	–	$8.53_{\pm 0.17}$	40.15	$17.34_{\pm 1.41}$	$28.46_{\pm 1.29}$	–	$8.26_{\pm 0.19}$	29.64
LR	$48.45_{\pm 0.16}$	$64.59_{\pm 0.19}$	$71.79_{\pm 0.12}$	$7.94_{\pm 0.08}$	16.27	$43.32_{\pm 0.14}$	$58.33_{\pm 0.17}$	$70.83_{\pm 0.13}$	$7.86_{\pm 0.05}$	10.03
ECC	$48.13_{\pm 0.18}$	$64.17_{\pm 0.13}$	$75.51_{\pm 0.17}$	$8.28_{\pm 0.11}$	15.81	$41.89_{\pm 0.19}$	$56.61_{\pm 0.12}$	$70.19_{\pm 0.16}$	$7.91_{\pm 0.09}$	9.48
RETAIN	$48.22_{\pm 0.15}$	$64.58_{\pm 0.17}$	$75.44_{\pm 0.11}$	$8.15_{\pm 0.06}$	18.95	$40.27_{\pm 0.13}$	$55.83_{\pm 0.19}$	$66.11_{\pm 0.12}$	$8.09_{\pm 0.08}$	10.93
LEAP	$46.28_{\pm 0.19}$	$63.90_{\pm 0.12}$	$74.56_{\pm 0.18}$	$7.56_{\pm 0.03}$	18.78	$40.09_{\pm 0.17}$	$55.17_{\pm 0.11}$	$58.80_{\pm 0.15}$	$7.24_{\pm 0.06}$	11.75
MICRON	$49.95_{\pm 0.11}$	$65.92_{\pm 0.16}$	$75.24_{\pm 0.13}$	$7.42_{\pm 0.09}$	18.64	$43.64_{\pm 0.15}$	$58.16_{\pm 0.18}$	$67.26_{\pm 0.12}$	$7.19_{\pm 0.07}$	13.31
GAMENet	$49.14_{\pm 0.14}$	$65.03_{\pm 0.18}$	$73.98_{\pm 0.12}$	$8.01_{\pm 0.05}$	24.36	$43.05_{\pm 0.20}$	$57.97_{\pm 0.13}$	$67.54_{\pm 0.17}$	$7.92_{\pm 0.09}$	15.26
VITA	$52.76_{\pm 0.16}$	$67.94_{\pm 0.13}$	$76.38_{\pm 0.11}$	$7.46_{\pm 0.08}$	20.64	$47.06_{\pm 0.14}$	$61.92_{\pm 0.12}$	$69.85_{\pm 0.14}$	$7.12_{\pm 0.05}$	12.55
SafeDrug	$50.62_{\pm 0.17}$	$66.71_{\pm 0.11}$	$74.96_{\pm 0.16}$	$6.75_{\pm 0.08}$	19.23	$44.82_{\pm 0.18}$	$60.06_{\pm 0.15}$	$68.87_{\pm 0.20}$	$6.69_{\pm 0.04}$	12.38
MoleRec	$52.55_{\pm 0.13}$	$68.00_{\pm 0.20}$	$77.12_{\pm 0.14}$	$7.41_{\pm 0.07}$	20.78	$46.23_{\pm 0.11}$	$61.21_{\pm 0.16}$	$69.16_{\pm 0.19}$	$6.93_{\pm 0.02}$	12.42
DEPOT	$52.92_{\pm 0.16}$	$68.78_{\pm 0.12}$	$77.64_{\pm 0.19}$	$7.32_{\pm 0.05}$	20.37	$46.81_{\pm 0.13}$	$61.75_{\pm 0.20}$	$69.98_{\pm 0.11}$	$6.81_{\pm 0.08}$	12.37
MedAlign	$53.90_{\pm 0.18}$	$69.16_{\pm 0.14}$	$78.14_{\pm 0.11}$	$7.29_{\pm 0.04}$	20.43	$47.94_{\pm 0.16}$	$62.54_{\pm 0.12}$	$70.61_{\pm 0.15}$	$6.74_{\pm 0.07}$	11.20
`GiantMed` $_{SD}$	$53.26_{\pm 0.15}$	$68.71_{\pm 0.19}$	$75.50_{\pm 0.13}$	$\mathbf{6.52}_{\pm 0.09}$	18.09	$48.70_{\pm 0.17}$	$63.77_{\pm 0.14}$	$69.25_{\pm 0.18}$	$\mathbf{6.37}_{\pm 0.05}$	11.51
`GiantMed` $_{MR}$	$54.84_{\pm 0.12}$	$69.97_{\pm 0.16}$	$77.42_{\pm 0.18}$	$7.38_{\pm 0.07}$	20.25	$49.52_{\pm 0.19}$	$64.54_{\pm 0.11}$	$70.65_{\pm 0.15}$	$6.74_{\pm 0.03}$	12.74
`GiantMed` $_{DP}$	$55.96_{\pm 0.14}$	$70.90_{\pm 0.11}$	$78.01_{\pm 0.15}$	$7.26_{\pm 0.02}$	19.93	$50.01_{\pm 0.13}$	$64.96_{\pm 0.15}$	$71.46_{\pm 0.12}$	$6.72_{\pm 0.04}$	11.05
`GiantMed` $_{MA}$	$\mathbf{56.63}_{\pm 0.13}$	$\mathbf{71.46}_{\pm 0.15}$	$\mathbf{78.48}_{\pm 0.11}$	$7.19_{\pm 0.04}$	19.99	$\mathbf{50.69}_{\pm 0.12}$	$\mathbf{65.73}_{\pm 0.18}$	$\mathbf{72.57}_{\pm 0.14}$	$6.58_{\pm 0.06}$	12.02

Table 3: Ablation results (%) of our proposed `GiantMed` with different backbone LLMs on MIMIC-III.

Method	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
	`GiantMed` $_{SD}$			
Qwen3-8B	$\mathbf{53.26}_{\pm 0.15}$	$\mathbf{68.71}_{\pm 0.19}$	$\mathbf{75.50}_{\pm 0.13}$	$\mathbf{6.52}_{\pm 0.09}$
LLaMA3.1	$52.18_{\pm 0.15}$	$67.25_{\pm 0.12}$	$74.73_{\pm 0.19}$	$7.05_{\pm 0.04}$
Aloe-Beta	$51.63_{\pm 0.18}$	$67.67_{\pm 0.13}$	$74.40_{\pm 0.16}$	$6.94_{\pm 0.11}$
	`GiantMed` $_{MA}$			
Qwen3-8B	$\mathbf{56.63}_{\pm 0.13}$	$\mathbf{71.46}_{\pm 0.15}$	$\mathbf{78.48}_{\pm 0.11}$	$\mathbf{7.19}_{\pm 0.04}$
LLaMA3.1	$54.98_{\pm 0.16}$	$69.89_{\pm 0.19}$	$75.38_{\pm 0.12}$	$7.32_{\pm 0.07}$
Aloe-Beta	$55.08_{\pm 0.14}$	$69.93_{\pm 0.11}$	$75.59_{\pm 0.18}$	$7.26_{\pm 0.03}$



(a) GiantMed$_{SD}$ **(b) GiantMed$_{MA}$**

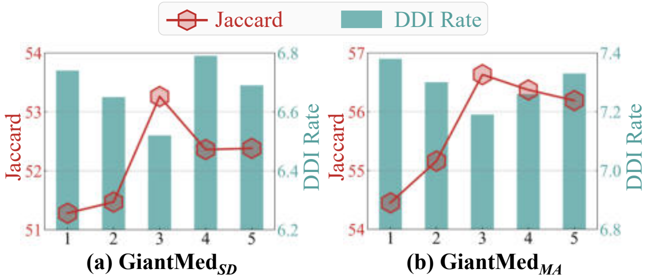
Figure 3: Hyperparameter results (%) of our `GiantMed` with varying boundary region width ϵ on MIMIC-III.

average of 19.79, while incurring a high DDI rate of 8.67%. In contrast, `GiantMed` leverages deep model guidance to direct LLMs' attention to clinically ambiguous boundary medications with fine-grained probabilities, yielding performance gains of up to 187.32% in both precision and safety metrics.

Compared with instance-based methods. Instance-based models, such as LR and ECC, rely solely on static features from individual visits. Their limited ability to capture temporal and contextual dependencies often leads to fewer medications being recommended (e.g., only 16.27 by LR on MIMIC-III), which in turn leads to lower overall precision. By contrast, `GiantMed` achieves a better balance between recommendation coverage and accuracy.

Compared with longitudinal methods. While longitudinal models like SafeDrug and MedAlign leverage sequential information to model patient history, they still struggle with capturing the contextual reasoning necessary for handling complex clinical scenarios, especially for medications near the decision boundary. Compared to these deep models, `GiantMed` integrates their accurate predictive control with the semantic reasoning capabilities of LLMs. This hybrid strategy enables `GiantMed` to achieve superior performance in both accuracy and safety across all evaluation settings.

Furthermore, each variant of our `GiantMed` is guided by a specific deep model (i.e., GiantMed$_{SD}$, GiantMed$_{MR}$, GiantMed$_{DP}$, and GiantMed$_{MA}$ guided by SafeDrug, MoleRec, DEPOT, and MedAlign, respectively). When directly compared with their corresponding deep model baselines, these variants achieve consistent and substantial performance gains, improving by up to 11.87% in Jaccard over SafeDrug and 5.06% in Jaccard over MedAlign. These results demonstrate the effectiveness of our boundary-aware LLM activation mechanism under deep model guidance. Notably, `GiantMed` focuses on leveraging fine-grained probabilistic outputs, rather than relying on any specific deep model, to identify clinically ambiguous medications, enabling flexible integration and strong adaptability across diverse clinical scenarios.



(a) GiantMed$_{SD}$ (b) GiantMed$_{MA}$

Figure 4: Hyperparameter results (%) of our proposed `GiantMed` with varying top-k retrieved similar EHRs on MIMIC-III.

Table 4: Ablation results (%) of our proposed `GiantMed` with different retrieval strategies on MIMIC-III.

Metric	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
Method	`GiantMed` $_{SD}$			
w. Diag.	**53.26**	**68.71**	**75.50**	**6.52**
w. Proc.	52.39	67.85	75.05	6.54
w. Diag. + Proc.	52.05	68.03	75.05	6.60
w/o. Retrieval	48.82	64.80	74.12	6.72
Method	`GiantMed` $_{MA}$			
w. Diag.	**56.63**	**71.46**	**78.48**	**7.19**
w. Proc.	55.90	70.61	76.17	7.30
w. Diag. + Proc.	56.03	70.85	76.28	7.33
w/o. Retrieval	53.28	68.51	75.73	7.26

5.3 Impact of Different Backbone LLMs (RQ2)

To evaluate the generalizability of `GiantMed` across different LLM backbones, we fix the underlying deep model and compare performance using three representative LLMs: the general-purpose Qwen3-8B and LLaMA3.1, and the domain-adapted Aloe-Beta. Table 3 reports the results of these LLMs when guided by either SafeDrug or MedAlign (Additional results with MoleRec and DEPOT are provided in Appendix B). Performance varies across LLM backbones under both SafeDrug and MedAlign guidance, with Qwen3-8B consistently achieving the best results across all metrics. For instance, with MedAlign guidance, Qwen3-8B outperforms Aloe-Beta by up to 3.82% in PRAUC, highlighting the advantage of its enhanced reasoning and instruction-following capabilities. These make it well-suited for complex clinical decision scenarios, particularly in refining boundary medications. Furthermore, the consistent trends observed across diverse LLMs show the flexibility and generalizability of our `GiantMed` framework.

5.4 Hyperparameter Sensitivity (RQ3)

We evaluate how two key hyperparameters (i.e., boundary region width ϵ and top-k retrieved similar EHRs) impact the performance and clarify how to set them (shown in Figure 3 and Figure 4).

Varying boundary region width ϵ. Figure 3 illustrates the impact of varying ϵ on Jaccard score and DDI rate for `GiantMed` $_{SD}$ and `GiantMed` $_{MA}$. As ϵ increases from 0.1 to 0.5, performance first improves and then declines, with the Jaccard score peaking at ϵ = 0.2 for both variants. A small ϵ (e.g., 0.1) yields too few boundary medications identified by deep models, limiting the benefits of semantic reasoning of LLMs. Conversely, a large ϵ (e.g., 0.5) overly expands the boundary region and includes many less ambiguous medications, weakening the refinement focus. We set $\epsilon = 0.2$ to balance the trade-off between refinement region and precision.

Table 5: Ablation results (%) of our proposed `GiantMed` with and without DDI constraints on MIMIC-III.

Metric	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
Method	`GiantMed` $_{SD}$			
w. DDI	53.26	68.71	**75.50**	**6.52**
w/o. DDI	**53.34**	**68.93**	75.36	6.94
Method	`GiantMed` $_{MA}$			
w. DDI	56.63	71.46	**78.48**	**7.19**
w/o. DDI	**56.85**	**71.99**	77.34	7.41

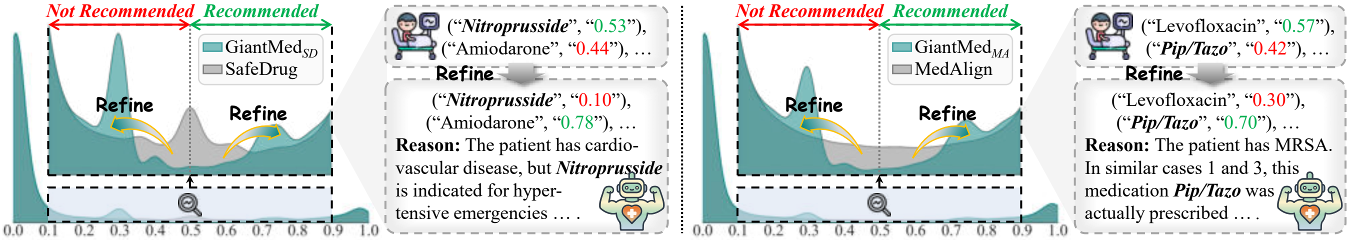
Varying top-k retrieved similar EHRs. To enrich semantic understanding and support accurate LLM reasoning over boundary medications, we retrieve the top-k most similar EHRs to supplement relevant clinical information. As shown in Figure 4, we observe that performance generally improves with increasing k, peaking at $k = 3$ for both GiantMed$_{SD}$ and GiantMed$_{MA}$. When k is too small, the LLM lacks sufficient contextual knowledge; when k is too large, irrelevant or redundant information may be introduced, slightly degrading performance. Therefore, we set $k = 3$ as the default, enabling LLMs to receive enough clinical context. More experimental results and analyses can be found in Appendix C.

5.5 Impact of Different Boundary Medication Augmentation Strategies (RQ4)

We investigate the impact of different boundary medication augmentation strategies in `GiantMed`, including retrieval-based clinical augmentation using similar EHRs and DDI constraint integration.

Impact of Different Retrieval Strategies. To identify the effective way to retrieve historical EHRs, we test four strategies: using diagnosis overlap (w. Diag.), procedure overlap (w. Proc.), both (w. Diag. + Proc.), and no retrieval (w/o. Retrieval). As shown in Table 4, all retrieval-based strategies consistently outperform the no-retrieval baseline across all `GiantMed` variants, confirming that external clinical context is critical for effective LLM refinement. Among them, retrieving based solely on diagnosis overlap (w. Diag.) consistently achieves the best results across all accuracy metrics and deep model variants. For instance, `GiantMed` $_{MA}$ reaches its highest Jaccard score of 56.63% with this strategy. In contrast, incorporating procedures either alone or in combination results in decreasing performance, likely due to procedure codes introducing more variability and less stable signals of clinical similarity. The substantial performance drop observed in the w/o. Retrieval setting further highlights that LLMs struggle to refine boundary medications without contextual support. Therefore, we adopt diagnosis-based retrieval as the default strategy for clinical augmentation. More experimental results and analyses can be found in Appendix D.

Impact of DDI Constraint Integration. To assess the impact of explicit pharmacological knowledge, we compare the performance of `GiantMed` with and without DDI constraint integration. As shown in Table 5, incorporating DDI constraints (w. DDI) consistently reduces the DDI rate across all `GiantMed` variants. For instance, with `GiantMed` $_{MA}$, the DDI rate drops from 7.41% to 7.19%, demonstrating improved medication safety. Moreover, the



(a) GiantMed$_{SD}$ vs. SafeDrug **(b) GiantMed$_{MA}$ vs. MedAlign**

Figure 5: Case studies on MIMIC-III, comparing medication probability distributions via kernel density estimation from deep models to our GiantMed, which further refines boundary medications through activating the reasoning potential of LLMs.

Table 6: Comparison of different methods on MIMIC-III, where "Avg. Runtime" denotes the average inference time.

Method	Jaccard (%) ↑	DDI (%) ↓	Avg. Runtime (s)
GiantMed $_{SD}$	$53.26_{\pm 0.15}$	$6.52_{\pm 0.09}$	12.47
GiantMed $_{MR}$	$54.84_{\pm 0.12}$	$7.38_{\pm 0.07}$	13.69
GiantMed $_{DP}$	$55.96_{\pm 0.14}$	$7.26_{\pm 0.02}$	12.49
GiantMed $_{MA}$	$56.63_{\pm 0.13}$	$7.19_{\pm 0.04}$	13.66
Qwen3-8B	$22.43_{\pm 2.05}$	$8.67_{\pm 0.24}$	284.26
Aloe-Beta	$22.33_{\pm 1.34}$	$8.53_{\pm 0.17}$	129.11
SafeDrug	$50.62_{\pm 0.17}$	$6.75_{\pm 0.08}$	0.23
MoleRec	$52.55_{\pm 0.13}$	$7.41_{\pm 0.07}$	0.32
DEPOT	$52.92_{\pm 0.16}$	$7.32_{\pm 0.05}$	0.36
MedAlign	$53.90_{\pm 0.18}$	$7.29_{\pm 0.04}$	0.45

slightly higher accuracy achieved without DDI constraints is expected due to non-negligible DDIs in the ground-truth prescriptions (e.g., 8.15% in MIMIC-III), reflecting a favorable safety–accuracy trade-off. By explicitly injecting DDI knowledge, we guide LLMs to avoid potentially harmful medication combinations while maintaining competitive recommendation performance. These results confirm that the DDI constraint integration is a vital component for generating clinically safe and reliable medication recommendations. More experimental results and analyses can be found in Appendix D.

5.6 Efficiency Analysis (RQ5)

As shown in Table 6, we evaluate the computational efficiency of GiantMed by measuring the average inference time on the MIMIC-III test set. While incorporating LLMs naturally incurs additional computational overhead compared to deep models, our GiantMed maintains a practical trade-off between performance and efficiency. Compared to LLM-based methods, which incur high inference costs (e.g., 284.26s for Qwen3-8B), our GiantMed is over 20× faster by activating LLMs only on small boundary medications. Although GiantMed is moderately slower than deep models (e.g., SafeDrug), it achieves significantly higher accuracy and safety, while providing enhanced interpretability through advanced semantic reasoning of LLMs (shown in Section 5.7). Notably, all GiantMed variants outperform their associated deep models with only a marginal increase in runtime, highlighting their clinical practicality.

5.7 Interpretable Case Studies (RQ6)

To qualitatively illustrate how GiantMed activates LLMs to enhance medication recommendations, we present two case studies from the MIMIC-III test set in Figure 5. These examples visualize how probability distributions are refined and highlight the clinical reasoning triggered by boundary-aware LLM activation.

From a distributional perspective, baseline deep models such as SafeDrug and MedAlign (gray curves) often concentrate probability mass near the decision boundary (e.g., around 0.5), indicating an ambiguous decision in medication selection. In contrast, GiantMed (teal curves) redistributes these probabilities toward the confident region (close to 0 or 1), thereby reducing clinical ambiguity and enhancing the accuracy in the final recommendation.

At the instance level, the case studies demonstrate how GiantMed applies fine-grained clinical reasoning to refine predictions. In Figure 5(a), it lowers the probability of "Nitroprusside" from 0.53 to 0.10 by recognizing that, despite the patient having cardiovascular disease, the medication is specific to hypertensive emergencies. Furthermore, in Figure 5(b), it raises the probability of "Pip/Tazo" from 0.42 to 0.70 by referencing similar EHRs where the medication was effectively used for MRSA infections, thus enabling context-aware, confident adjustments for appropriate prescription. These cases demonstrate GiantMed's ability to refine ambiguous boundary medications by leveraging pharmacological knowledge with augmented clinical context, effectively activating the reasoning capacity of LLMs for more accurate and interpretable decisions.

6 Conclusion

In this paper, we propose GiantMed, a novel deep model-guided LLM framework for boundary-aware medication recommendation. GiantMed first identifies boundary medications using fine-grained probability outputs from a trained deep model, then refines the prediction over these medications through a boundary-aware prompt enhanced with retrieved historical EHRs and integrated DDI constraints. Experimental results on MIMIC-III and MIMIC-IV show that GiantMed consistently outperforms state-of-the-art baselines in terms of both accuracy and safety. Further analyses demonstrate the flexibility and efficiency of our framework, highlighting its practical utility in real-world clinical settings.

Acknowledgements

This work was supported in part by the Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grant (2025H0042), Fujian Provincial Natural Science Foundation of China under Grants (2025J01540), National Natural Science Foundation of China under Grants (62302098). Carl Yang was not supported by any fund from China.

References

[1] P.L. Bartlett. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* 44, 2 (1998), 525–536.

[2] Peter L Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research* 3, Nov (2002), 463–482.

[3] Ofir Ben Shoham and Nadav Rappoport. 2024. Cpllm: Clinical prediction with large language models. *PLOS Digital Health* 3, 12 (2024), e0000680.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, Vol. 33. 1877–1901.

[5] Weiwei Cheng and Eyke Hüllermeier. 2009. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning* 76 (2009), 211–225.

[6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in neural information processing systems*, Vol. 29.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. In *Proceedings of the 41st International Conference on Machine Learning*.

[8] Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1543–1558.

[9] Chenxiao Fan, Chongming Gao, Wentao Shi, Yaxin Gong, Zihao Zhao, and Fuli Feng. 2025. Fine-grained List-wise Alignment for Generative Medication Recommendation. In *Advances in neural information processing systems*, Vol. 38.

[10] Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. 2025. The Aloe Family Recipe for Open and Specialized Healthcare LLMs. (2025). arXiv:2505.04388

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. 1321–1330.

[12] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion* 118 (2025), 102963.

[13] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2607–2615.

[14] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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

[16] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. 2018. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association* 25, 1 (2018), 32–39.

[17] Yan Kang, Mingjian Yang, Yue Peng, Jingwen Cai, Lei Zhao, Zhan Gao, Ningshu Li, and Bin Pu. 2025. LLM-DG: Leveraging large language model for enhanced disease prediction via inter-patient and intra-patient modeling. *Information Fusion* 121 (2025), 103145.

[18] 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 AAAI Conference on Artificial Intelligence*, Vol. 39. 11950–11958.

[19] 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 AAAI Conference on Artificial Intelligence*, Vol. 38. 8600–8607.

[20] Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1637–1645.

[21] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024. Large language model distilling medication recommendation model. *arXiv preprint arXiv:2402.02803* (2024).

[22] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (2022).

[23] Hang Lv, Zixuan Guo, Zijie Wu, Yanchao Tan, Guofang Ma, Zhigang Lin, Xiping Chen, Hong Cheng, and Carl Yang. 2025. MedAlign: Enhancing Combinatorial Medication Recommendation with Multi-modality Alignment. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6084–6092.

[24] OpenAI. 2025. *GPT-5 System Card*. Technical Report. https://openai.com/index/gpt-5-system-card Accessed: 2025.

[25] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning* 85 (2011), 333–359.

[26] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1126–1133.

[27] Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2022. Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. In *Advances in Neural Information Processing Systems*, Vol. 35. 27837–27849.

[28] Jie Tan, Yu Rong, Kangfei Zhao, Tian Bian, Tingyang Xu, Junzhou Huang, Hong Cheng, and Helen Meng. 2024. Natural Language-Assisted Multi-modal Medication Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2200–2209.

[29] Yunsen Tang, Ning Liu, Haitao Yuan, Yonghe Yan, Lei Liu, Weixing Tan, and Lizhen Cui. 2024. LAMRec: Label-aware Multi-view Drug Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2230–2239.

[30] Vladimir N Vapnik and A Ya Chervonenkis. 2015. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*. 11–30.

[31] Huimin Wang, Wai-Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023. CoAD: Automatic Diagnosis through Symptom and Disease Collaborative Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6348–6361.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, Vol. 35. 24824–24837.

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

[34] Jialun Wu, Xinyao Yu, Kai He, Zeyu Gao, and Tieliang Gong. 2024. PROMISE: A pre-trained knowledge-infused multimodal representation learning framework for medication recommendation. *Information Processing & Management* 61, 4 (2024), 103758.

[35] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM web conference 2022*. 935–945.

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[37] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change Matters: Medication Change Prediction with Recurrent Residual Networks. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*. 3728–3734.

[38] 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 *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*. 3735–3741.

[39] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM web conference 2023*. 4075–4085.

[40] Qinkai Yu, Mingyu Jin, Dong Shu, Chong Zhang, Wenyue Hua, Mengnan Du, and Yongfeng Zhang. 2025. Health-LLM: Personalized Retrieval-Augmented Disease Prediction System: Health-LLM. In *ACL Workshop NLP for Positive Impact*. 1.

[41] 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.

[42] 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* (2024).

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

Appendix

A Details of Datasets

To verify the effectiveness of the compared methods, we use two real-world Electronic Health Record (EHR) datasets, i.e., **MIMIC-III** [15] and **MIMIC-IV** [16]. The publicly available MIMIC-III dataset includes over forty thousand adult patients treated in critical care units at the Beth Israel Deaconess Medical Center from 2001 to 2012. In contrast, the MIMIC-IV dataset provides more comprehensive EHRs and clinical information for hospitalized patients spanning from 2008 to 2019. During data preprocessing, we represent diagnoses and procedures using codes from the ninth revision of the International Classification of Diseases (ICD-9). Medications are encoded using the third level of the Anatomical Therapeutic Chemical (ATC-3) classification system, where each code denotes a specific chemical, pharmacological, or therapeutic group. Notably, each ATC code maps to one or more medications, and each medication corresponds to an ATC code. To ensure sufficient clinical history, we retain only patients with at least two visits. Consistent with prior studies [23, 38, 39, 42], we consider potential Drug-Drug Interactions (DDIs) by extracting the most frequent top-K DDI types from DrugBank [33]. We set $K = 40$ for fair comparison.

B Impact of Different Backbone LLMs

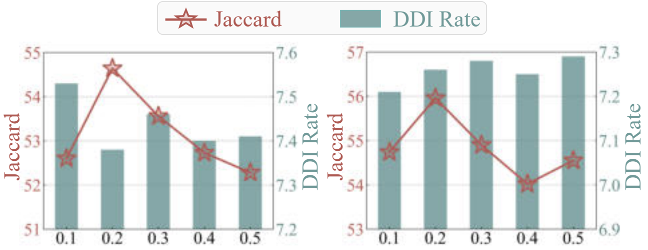
To evaluate the generalizability of `GiantMed` across different LLM backbones, we fix the underlying deep model and compare performance using three representative LLMs: the general-purpose Qwen3-8B and LLaMA3.1, and the domain-adapted Aloe-Beta. Table 7 reports the results of these LLMs when guided by either MoleRec or DEPOT. Performance varies across LLM backbones under both MoleRec and DEPOT guidance, with Qwen3-8B consistently achieving the best results across all metrics. For instance, with DEPOT guidance, Qwen3-8B outperforms Aloe-Beta by up to 2.73% in PRAUC, highlighting the advantage of its enhanced reasoning and instruction-following capabilities. These make it well-suited for complex clinical decision scenarios, particularly in refining boundary medications. Furthermore, the consistent trends observed across diverse LLMs show the flexibility and generalizability of our `GiantMed` framework.

Table 7: Ablation results (%) of our proposed `GiantMed` with different backbone LLMs on MIMIC-III.

Method	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
		`GiantMed` $_{MR}$		
Qwen3-8B	**54.84**$_{\pm0.12}$	**69.97**$_{\pm0.16}$	**77.42**$_{\pm0.18}$	**7.38**$_{\pm0.07}$
LLaMA3.1	53.04$_{\pm0.18}$	68.37$_{\pm0.14}$	75.21$_{\pm0.19}$	7.51$_{\pm0.04}$
Aloe-Beta	54.07$_{\pm0.16}$	69.38$_{\pm0.11}$	75.29$_{\pm0.18}$	7.56$_{\pm0.06}$
		`GiantMed` $_{DP}$		
Qwen3-8B	**55.96**$_{\pm0.14}$	**70.90**$_{\pm0.11}$	**78.01**$_{\pm0.15}$	**7.26**$_{\pm0.02}$
LLaMA3.1	54.29$_{\pm0.17}$	69.51$_{\pm0.13}$	75.37$_{\pm0.19}$	7.35$_{\pm0.11}$
Aloe-Beta	54.64$_{\pm0.15}$	69.90$_{\pm0.18}$	75.94$_{\pm0.12}$	7.29$_{\pm0.06}$

C Hyperparameter Sensitivity

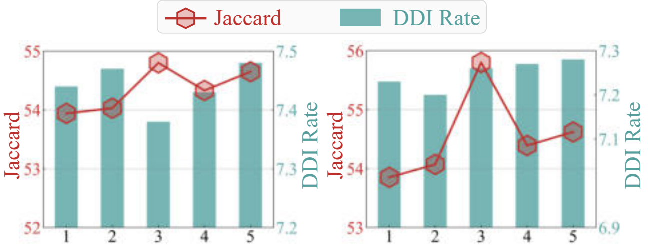
Varying boundary region width ϵ. Figure 6 illustrates the impact of varying ϵ on Jaccard score and DDI rate for `GiantMed` $_{MR}$ and `GiantMed` $_{DP}$. As ϵ increases from 0.1 to 0.5, performance first improves and then declines, with the Jaccard score peaking at $\epsilon = 0.2$ for both variants. A small ϵ (e.g., 0.1) yields too few boundary medications identified by deep models, limiting the benefits of semantic reasoning of LLMs. Conversely, a large ϵ (e.g., 0.5) overly expands the boundary region and includes many less ambiguous medications, weakening the refinement focus. We set $\epsilon = 0.2$ to balance the trade-off between refinement scope and precision.



(a) GiantMed$_{MR}$ **(b) GiantMed$_{DP}$**

Figure 6: Hyperparameter results (%) of our `GiantMed` with varying boundary region width ϵ on MIMIC-III.

Varying top-k retrieved similar EHRs. To enrich semantic understanding and support accurate LLM reasoning over boundary medications, we retrieve the top-k most similar EHRs to supplement relevant clinical information. As shown in Figure 7, we observe that performance generally improves with increasing k, peaking at $k = 3$ for both `GiantMed` $_{MR}$ and `GiantMed` $_{DP}$. When k is too small, the LLM lacks sufficient contextual knowledge; when k is too large, irrelevant or redundant information may be introduced, slightly degrading performance. Therefore, we set $k = 3$ as the default, enabling LLMs to receive enough clinical context.



(a) GiantMed$_{MR}$ **(b) GiantMed$_{DP}$**

Figure 7: Hyperparameter results (%) of our `GiantMed` with varying top-k retrieved similar EHRs on MIMIC-III.

D Impact of Different Boundary Medication Augmentation Strategies

Impact of Different Retrieval Strategies. To identify the effective way to retrieve historical EHRs, we test four strategies: using diagnosis overlap (w. Diag.), procedure overlap (w. Proc.), both (w. Diag. + Proc.), and no retrieval (w/o. Retrieval). As shown in Table 8, all retrieval-based strategies consistently outperform the no-retrieval baseline across all `GiantMed` variants, confirming that external clinical context is critical for LLM refinement. Among them, retrieving based solely on diagnosis overlap (w. Diag.) consistently achieves the best results across all accuracy metrics and deep model variants. For instance, `GiantMed` $_{DP}$ reaches its highest Jaccard score of 55.96% with this strategy. In contrast, incorporating procedures either alone or in combination results in decreasing performance, likely due to procedure codes introducing more variability and less stable signals of clinical similarity. The substantial performance drop observed in the w/o. Retrieval setting further highlights that LLMs struggle to refine boundary medications without contextual support. Therefore, we adopt diagnosis-based retrieval as the default strategy for clinical augmentation.

Table 8: Ablation results (%) of GiantMed with different retrieval strategies on MIMIC-III.

Metric	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
Method	GiantMed $_{MR}$			
w. Diag.	**54.84**	**69.97**	**77.42**	7.38
w. Proc.	53.89	69.16	75.32	7.58
w. Diag. + Proc.	54.28	69.47	75.35	7.46
w/o. Retrieval	51.26	66.82	74.87	**7.29**
Method	GiantMed $_{DP}$			
w. Diag.	**55.96**	**70.90**	**78.01**	7.26
w. Proc.	54.60	69.67	75.86	7.32
w. Diag. + Proc.	55.23	70.29	75.83	**7.18**
w/o. Retrieval	51.56	66.98	75.14	7.29

Impact of DDI Constraint Integration. To assess the impact of explicit pharmacological knowledge, we compare the performance of GiantMed with and without DDI constraint integration. As shown in Table 9, incorporating DDI constraints (w. DDI) consistently reduces the DDI rate across all GiantMed variants. For instance, with GiantMed $_{DP}$, the DDI rate drops from 7.53% to 7.26%, demonstrating improved medication safety. These results confirm that DDI constraint integration is a vital component for generating clinically safe and reliable medication recommendations.

Table 9: Ablation results (%) of GiantMed with and without DDI constraints on MIMIC-III.

Metric	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓
Method	GiantMed $_{MR}$			
w. DDI	**54.84**	**69.97**	**77.42**	**7.38**
w/o. DDI	54.57	68.82	77.04	7.54
Method	GiantMed $_{DP}$			
w. DDI	55.96	**70.90**	**78.01**	**7.26**
w/o. DDI	**56.03**	70.86	77.14	7.53

E Calibration Evaluation for Boundary Medication Prediction

We introduce the Expected Calibration Error (ECE) [11], which quantifies the difference between predicted confidence and actual accuracy. The metric is computed as: ECE = $\frac{1}{|M_{bound}|}\sum_{i\in M_{bound}}|\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i|$, where $|M_{bound}|$ denotes the number of boundary medications. Figure 8 shows that GiantMed achieves up to a 50.27% reduction in ECE, reflecting better boundary correction and higher reliability.

F Stability of LLM Refinement under Different Rephrased Prompts

To assess the stability of LLM refinements, we generate three rephrased versions of the refinement prompt via GPT-5, following [8]. Specifically, we apply the instruction as follows:

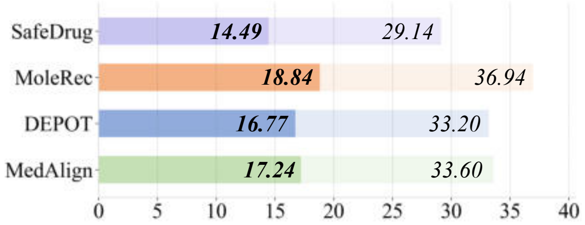


Figure 8: ECE results (%) on MIMIC-III for boundary medication prediction. Darker bars denote GiantMed, while lighter bars represent the corresponding baseline methods. Lower ECE indicates better calibration and reliability.

> **Instruction:** *Please rephrase the prompt into three versions, keeping the task unchanged while allowing mild operations such as synonym replacement and word reordering in task-unrelated parts.*

Table 10 shows that GiantMed remains consistently stable and robust across all prompt variants for GiantMed $_{MA}$ on the MIMIC-III dataset, maintaining superior performance over the corresponding deep model MedAlign.

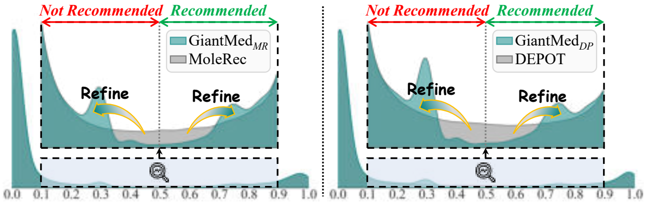
Table 10: Performance results (%) of GiantMed $_{MA}$ on MIMIC-III using three rephrased versions of the refinement prompt.

Method	Jaccard ↑	DDI ↓
MedAlign	$53.90_{\pm 0.18}$	$7.29_{\pm 0.04}$
GiantMed $_{MA}$ (Ours)	$56.63_{\pm 0.13}$	$7.19_{\pm 0.04}$
GiantMed $_{MA}$ (Rephrased 1)	$55.99_{\pm 0.12}$	$7.26_{\pm 0.07}$
GiantMed $_{MA}$ (Rephrased 2)	$56.01_{\pm 0.15}$	$7.21_{\pm 0.07}$
GiantMed $_{MA}$ (Rephrased 3)	$\mathbf{56.83_{\pm 0.12}}$	$\mathbf{7.18_{\pm 0.05}}$

G Interpretable Case Studies

To qualitatively illustrate how GiantMed activates LLMs to enhance medication recommendations, we present two case studies from the MIMIC-III test set in Figure 9. These examples visualize how probability distributions are refined and highlight the clinical reasoning triggered by boundary-aware LLM activation.

From a distributional perspective, baseline deep models such as MoleRec and DEPOT (gray curves) often concentrate probability mass near the decision boundary (e.g., around 0.5), indicating an ambiguous decision in medication selection. In contrast, GiantMed (teal curves) redistributes these probabilities toward the confident region (close to 0 or 1), thereby reducing clinical ambiguity and enhancing the accuracy in the final recommendation.



(a) GiantMed$_{MR}$ vs. MoleRec **(b) GiantMed$_{DP}$ vs. DEPOT**

Figure 9: Comparison of medication probability distributions on MIMIC-III from deep models to our GiantMed.