HEM-Med: A Hierarchical Agentic Memory Framework for Safe Multi-Center Medication Recommendation

Min Xiao[1,†], Jin Li[2,†], Shoujin Wang[2], Yeguang Zhen[1], Yishuo Li[1], Wenpeng Lu[1,*]
[1] 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
Email: min.xiao02@foxmail.com; zhenyeguang92@gmail.com;
yishuoli@foxmail.com; wenpeng.lu@qlu.edu.cn
[2] University of Technology Sydney
Email: Jin.Li-4@student.uts.edu.au; shoujinwang@foxmail.com
[†] Equal Contributors [*] Corresponding author

arXiv:2610.20191v1 [cs.AI] 14 Oct 2026

***Abstract*—Medication recommendation (MR) is critical for clinical decision support. While existing MR methods are mostly designed for single-center learning, where only one independent hospital is involved, multi-center MR with hospitals of different scales has become increasingly demanding in practice. In multi-center settings, abundant data from large medical centers can improve recommendation performance for low-resource centers. Despite a few preliminary attempts on this task, two critical gaps remain. Gap 1: Existing multi-center methods fail to disentangle transferable cross-hospital clinical experience from hospital-specific prescribing patterns, limiting clinical feasibility and physician acceptance. Gap 2: They commonly overlook multi-label safety constraints of MR, including drug-drug interaction risks and local formulary feasibility, which are essential for safe real-world deployment. To address these gaps, we propose HEM-Med, a novel hierarchical agentic memory framework for multi-center MR. To address Gap 1, we design a disentangled hierarchical memory module that builds patient-level medication-context attribution records and disentangles them into cross-hospital experience memory and hospital-specific residual memory. To address Gap 2, we introduce a safety-aware medication set generation module, which integrates dual-level experience retrieval with safety-aware set decoding to produce clinically relevant, safer, and locally adaptive medication recommendations. Experiments on a real-world multi-center eICU dataset show that HEM-Med outperforms representative transfer learning, LLM-based, and MR baselines in terms of both recommendation accuracy and safety.** [1]

***Index Terms*—Multi-center Medication Recommendation, Experience Memory, Drug-Drug Interaction, Memory Disentanglement**

I. Introduction

Medication recommendation (MR) is a key task in clinical decision support systems. The goal is to analyze a patient's complex Electronic Health Record (EHR) and generate a set of candidate medications that are well aligned with the patient's current physiological status and therapeutic needs [1], [2].

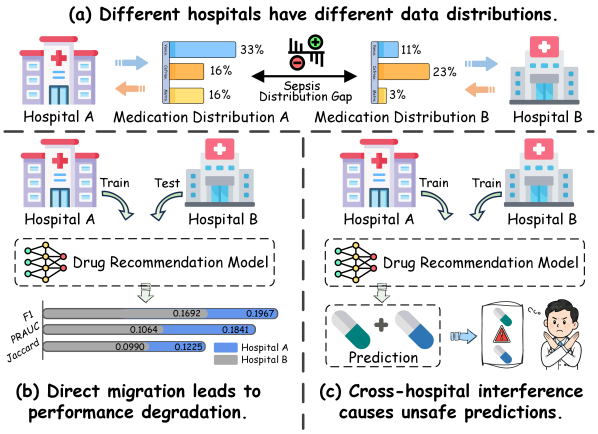


Fig. 1: Motivating example of multi-center medication recommendation. Heterogeneous hospital prescribing patterns may cause negative transfer, and ignoring local feasibility and drug safety may lead to unsafe medication combinations.

Early MR studies focused on model development within a single center, where training and evaluation are conducted using EHR data from the same hospital. Representative methods include GAMENet, SafeDrug, and DKGM [1]–[3]. Despite their effectiveness in single-center settings, these methods rely on local data and may perform poorly in low-resource hospitals owing to data sparsity and cold-start issues [4].

Recent studies address single-center limitations through two lines of research: cross-domain and multi-center transfer for cross-hospital knowledge sharing. The first line borrows ideas from General Cross-Domain Recommendation (CDR), which transfers or shares knowledge across related domains, such

[1]Our code is publicly available at: https://anonymous.4open.science/r/anonymous-HEM-Med-0935.

as EMCDR and DDTCDR [5], [6]. The second line targets multi-center healthcare scenarios. A representative method TEMPT [7] pre-trains on pooled multi-center prescription data and then relies solely on prompt fine-tuning to adapt to the target hospital distribution.

Despite recent progress, existing studies still suffer from two major gaps. **Gap 1: Entangled cross-hospital knowledge.** Existing multi-center MR methods often overlook that prescribing decisions reflect both generalizable medical knowledge and hospital-specific resources, clinical protocols, and prescribing preferences. As shown in Fig. 1 (a), medication distributions may vary across hospitals for the same disease [7]. Fig. 1 (b) further illustrates that transferring entangled cross-hospital knowledge in ICU settings with coexisting conditions may cause negative transfer and impair recommendation performance. **Gap 2: Insufficient safety awareness.** Existing transfer-based methods prioritize representation alignment but often overlook clinical safety constraints. Unlike general CDR, MR requires multi-label predictions that satisfy treatment needs, DDI constraints, and local formulary availability [2], [8], [9]. As illustrated in Fig. 1 (c), ignoring these constraints may yield inappropriate or unsafe medication combinations.

To address these gaps, we propose HEM-Med, a hierarchical agentic memory framework for multi-center MR. For Gap 1, a disentangled memory module separates transferable clinical experience from hospital-specific prescribing patterns into cross-hospital experience memory and hospital-specific residual memory. For Gap 2, a safety-aware medication set generation module combines dual-level retrieval with DDI-aware set decoding to generate clinically relevant, safer, and locally adaptive recommendations.

The contributions are summarized as follows:

- We propose HEM-Med, a hierarchical agentic memory framework for multi-center MR that explicitly disentangles cross-hospital experience from hospital-specific prescribing preferences.
- We develop a safety-aware module for ICU medication set generation, integrating dual-level retrieval and DDI-aware decoding to account for clinical relevance, local feasibility, and DDI risks.
- Experiments on the eICU dataset show that HEM-Med improves recommendation accuracy while maintaining competitive medication safety.

II. Related Work

A. Medication Recommendation

EHR-based MR aims to recommend personalized, clinically safe medication combinations. Existing approaches include longitudinal modeling with RETAIN and DMNC [10], [11], time-aware modeling with MR-DTR [12], and DDI-aware recommendation with GAMENet, SafeDrug, and MoleRec [1], [2], [13]. DKINet, TCMPR, and SECGI [14]–[16] enhance patient representations with external knowledge or semantics. However, most methods target single-center settings and may suffer from negative transfer across hospitals.

B. Cross-Domain Transfer and Multi-Center Learning

CDR and multi-center transfer alleviate data sparsity through knowledge sharing. EMCDR [5] and DisenCDR [17] align domain representations, while TEMPT [7] adapts to hospitals through pre-training and prompt-tuning. However, general CDR overlooks combination therapy and DDI risks, while medical transfer lacks explicit consensus–preference disentanglement, risking negative transfer.

C. LLM Agents and Knowledge-Augmented Healthcare

LLMs support clinical decision-making through semantic reasoning [18]. Knowledge-enhanced methods include multi-agent collaboration (MedAgents, MDTeamGPT [19], [20]) and retrieval- or tool-augmented reasoning (ClinicalGPT, TxAgent, JointCoder, SOAPTriage [21]–[24]). However, their prompt-based knowledge integration often lacks explicit separation of general medical safety principles from hospital-specific execution rules for multi-center MR.

III. Preliminaries

A. Multi-Center Electronic Health Records Data

Multi-center EHR data comprise independent ICU records from multiple hospitals. Let $\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$ denote the hospital set, and $\mathcal{D}$, $\mathcal{P}$, $\mathcal{O}$, and $\mathcal{M}$ the diagnosis, procedure, abnormal laboratory signal, and standardized medication vocabularies, respectively. The i-th record is represented as:

$$\mathbf{r}_i = (h_i, D_i, P_i, O_i, M_i), \tag{1}$$

where $h_i \in \mathcal{H}$ is the hospital identifier, $D_i \subseteq \mathcal{D}$, $P_i \subseteq \mathcal{P}$, $O_i \subseteq \mathcal{O}$, and $M_i \subseteq \mathcal{M}$ denote the observed diagnoses, procedures, abnormal laboratory signals, and medications, respectively.

B. Clinical Anchors

The clinical anchor set for EHR instance i comprises observed diagnoses, procedures, and abnormal laboratory signals:

$$\begin{aligned} A_i = &\{(\text{diag}, d) \mid d \in D_i\} \cup \{(\text{proc}, p) \mid p \in P_i\} \\ &\cup \{(\text{lab}, o) \mid o \in O_i\}. \end{aligned} \tag{2}$$

Each anchor $a \in A_i$ represents a typed clinical signal: a diagnosis, procedure, or abnormal laboratory state.

C. Drug-Drug Interaction Knowledge

We denote the drug-drug interaction knowledge as a graph:

$$\Omega_{\text{DDI}} = (\mathcal{M}, \mathcal{E}_{\text{DDI}}), \tag{3}$$

where $\mathcal{M}$ is the standardized medication vocabulary and $\mathcal{E}_{\text{DDI}} \subseteq \mathcal{M} \times \mathcal{M}$ contains potentially interacting medication pairs. Each node represents a medication, and each edge (u, v) represents a potential interaction with severity score $\text{sev}(u, v)$.

D. Medication Recommendation

Based on the observed clinical context of the i-th EHR instance, we define the model input as:

$$\mathbf{x}_i = (h_i, D_i, P_i, O_i), \tag{4}$$

while M_i is the ground-truth set. Multi-center MR estimates conditional relevance scores $s(m \mid \mathbf{x}_i)$ for $m \in \mathcal{M}$ and decodes the medication set $\hat{Y}_i \subseteq \mathcal{M}$ as:

$$\hat{Y}_i = \mathrm{Decode}\left(\{s(m \mid \mathbf{x}_i)\}_{m\in\mathcal{M}}, \Omega_{\mathrm{DDI}}, \mathbf{x}_i\right). \tag{5}$$

IV. Method

We propose HEM-Med, a hierarchical agentic memory framework for multi-center MR. As shown in Fig. 2, it comprises two components: (1) consensus–residual memory disentanglement, which separates historical medication–context experience into cross-hospital consensus and hospital-specific residual memories; and (2) safety-aware recommendation, which combines dual-level retrieval, experience-aware reranking, DDI-aware set decoding, and safety review to generate locally feasible and clinically safe medication sets.

A. Patient-Level Attribution Memory

Raw ICU EHRs lack explicit medication links to diagnoses, procedures, or laboratory abnormalities. HEM-Med builds auditable L_3 patient-level memory via constrained DeepSeek-assisted attribution.

For each training ICU stay s_i and each observed medication $m_j \in M_i$, we construct a stay-medication query:

$$q_{ij} = (h_i, m_j, D_i, P_i, O_i), \tag{6}$$

where h_i denotes the hospital identifier, and D_i, P_i, and O_i denote the observed diagnoses, procedures, and abnormal laboratory signals of the current stay, respectively.

DeepSeek-v4-flash selects at most one observed context anchor for constrained attribution, without adding external indications or clinical facts:

$$\begin{aligned} a_{ij} \in &\{\mathrm{diag}(d) \mid d \in D_i\} \cup \{\mathrm{proc}(p) \mid p \in P_i\} \\ &\cup \{\mathrm{lab}(o) \mid o \in O_i\} \cup \{\emptyset\}. \end{aligned} \tag{7}$$

Insufficient context yields $\emptyset$. Attribution is a conservative empirical hypothesis, not a verified prescribing intention, indication, or causal treatment relation.

For each validated stay-medication pair, the L_3 record is defined as:

$$e_{ij}^{(L_3)} = (s_i, h_i, m_j, a_{ij}, r_{ij}, S_{ij}), \tag{8}$$

where r_{ij} denotes the conservative attribution hypothesis, and S_{ij} stores training-set statistical support for subsequent aggregation.

Training-only support statistics distinguish cross-hospital experience from hospital-specific residual patterns in anchor–medication associations.

All validated L_3 training records are collected as:

$$\mathcal{K}^{(L_3)} = \left\{ e_{ij}^{(L_3)} \mid s_i \in \mathcal{S}_{\mathrm{train}}, m_j \in M_i \right\}. \tag{9}$$

Three clinicians reviewed 300 randomly sampled L_3 records, yielding Fleiss' κ values of 0.813 for anchor selection and 0.726 for attribution assessment.

B. Consensus and Residual Memory Disentanglement

HEM-Med separates patient-level records into cross-hospital experience and hospital-specific residual memories.

Cross-hospital experience memory. Let $n(a)$ count occurrences of clinical anchor a, $n(a,m)$ count its co-occurrences with medication m, and $H(a,m)$ count hospitals observing (a,m). The anchor-conditioned medication frequency is:

$$p(m \mid a) = \frac{n(a,m)}{n(a)}. \tag{10}$$

The cross-hospital consensus memory $\mathcal{K}^{(L_1)}$ stores associations that satisfy sufficient support, frequency, and hospital-coverage constraints:

$$\mathcal{K}^{(L_1)} = \{(a,m) \mid \mathbf{c}(a,m) \succeq \boldsymbol{\tau},\ p(m \mid a) \geq \tau_p\}. \tag{11}$$

Here, $\mathbf{c}(a,m) = [n(a), n(a,m), H(a), H(a,m)]$ and $\boldsymbol{\tau} = [\tau_a, \tau_{am}, \tau_{ha}, \tau_h]$. $H(a)$ denotes the hospital coverage of anchor a, and $\succeq$ denotes element-wise comparison.

Hospital-specific residual memory. For hospital h, anchor a, and medication m, we define:

$$p_h(m \mid a) = \frac{n_h(a,m)}{n_h(a)}, \tag{12}$$

where $n_h(a)$ and $n_h(a,m)$ denote the support and co-occurrence counts within hospital h, respectively. The residual deviation is computed as:

$$\Delta_h(a,m) = p_h(m \mid a) - p(m \mid a). \tag{13}$$

Let $\tau_a^{(L_2)}$, $\tau_{am}^{(L_2)}$, $\tau_p^{(L_2)}$, and $\tau_\Delta^{(L_2)}$ be predefined thresholds. Only local associations with sufficient hospital-level support and residual magnitude are retained in L_2 memory:

$$\begin{aligned} \mathcal{K}^{(L_2)} = \Big\{ (h,a,m,\Delta_h(a,m), c_{h,a,m}^{(L_2)}) \ \Big|\ & n_h(a) \geq \tau_a^{(L_2)}, \\ & n_h(a,m) \geq \tau_{am}^{(L_2)}, p_h(m \mid a) \geq \tau_p^{(L_2)}, \\ & |\Delta_h(a,m)| \geq \tau_\Delta^{(L_2)} \Big\}. \end{aligned} \tag{14}$$

Here, $c_{h,a,m}^{(L_2)}$ denotes the residual type: amplified, suppressed, or localized. L_2 complements L_1 by retaining hospital-specific residual prescribing patterns.

Weak L_1 and Weak L_2 Experience. HEM-Med aggregates and down-weights medication–context associations below strong L_1 or L_2 thresholds as weak experience to support inference recall for long-tail medications and rare clinical contexts.

Associations with cross-hospital support below strict L_1 thresholds are retained as weak L_1 experience:

$$K_{\mathrm{weak}}^{(L_1)} = \left\{ (a,m) \mid (a,m) \notin \mathcal{K}^{(L_1)},\ \mathcal{C}_{\mathrm{weak}}^{(L_1)}(a,m) = 1 \right\}, \tag{15}$$

where $\mathcal{C}_{\mathrm{weak}}^{(L_1)}(a,m)$ denotes relaxed cross-hospital criteria based on counts, conditional frequency, and hospital coverage.

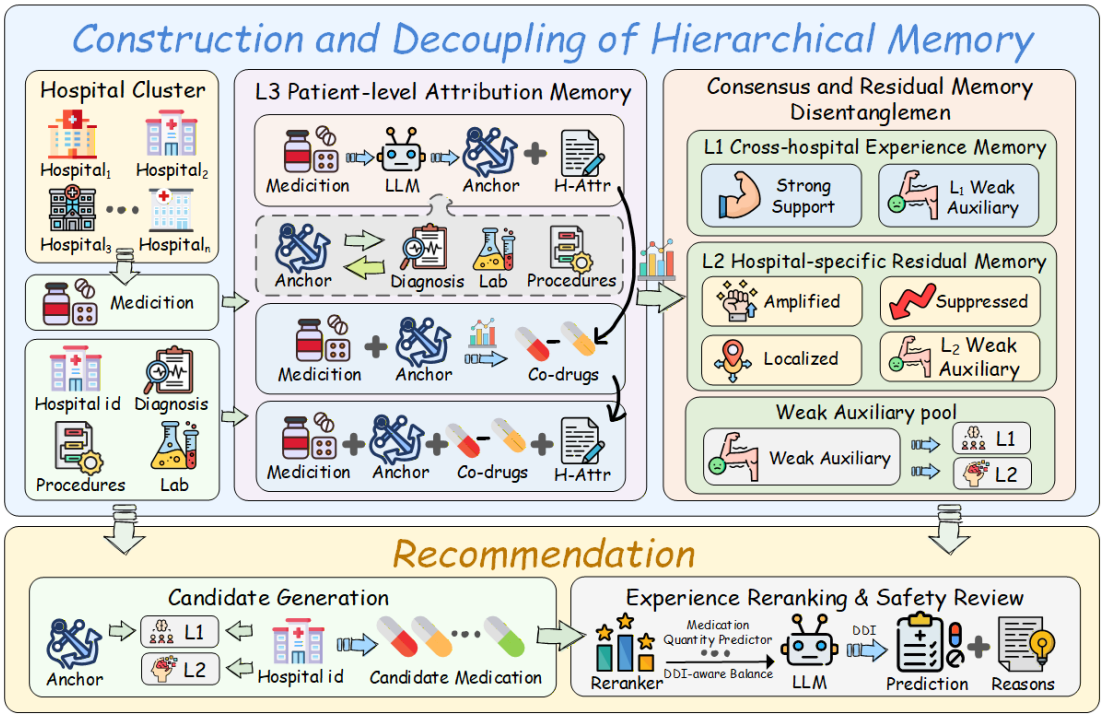


Fig. 2: Overall agentic framework of HEM-Med. HEM-Med retrieves disentangled cross-hospital and hospital-specific experience, and generates medication sets through calibrated prediction and DDI-aware decoding.

Similarly, hospital-specific associations with sufficient local support but failing strict L_2 residual criteria are retained as weak L_2 experience:

$$K^{(L_2)}_{\text{weak}} = \{(h, a, m, \Delta_h(a, m), \rho^{(L_2)}_{h,a,m}) \mid (h, a, m) \notin \mathcal{K}^{(L_2)}, \mathcal{C}^{(L_2)}_{\text{weak}}(h, a, m) = 1\}, \tag{16}$$

where $\mathcal{C}^{(L_2)}_{\text{weak}}(h, a, m)$ defines relaxed local criteria for support, conditional frequency, and residual deviation.

During retrieval, weak experience is down-weighted and cannot dominate strong L_1 or L_2 experience:

$$g_{L_1}(m, \mathcal{A}_i) = g^{\text{strong}}_{L_1}(m, \mathcal{A}_i) + \gamma^{\text{weak}}_{L_1} g^{\text{weak}}_{L_1}(m, \mathcal{A}_i), \tag{17}$$

$$g_{L_2}(m, \mathcal{A}_i, h_i) = g^{\text{strong}}_{L_2}(m, \mathcal{A}_i, h_i) + \gamma^{\text{weak}}_{L_2} g^{\text{weak}}_{L_2}(m, \mathcal{A}_i, h_i), \tag{18}$$

where $g^{\text{strong}} L_k$ and $g^{\text{weak}} L_k$ denote strong and weak experience scores, respectively, and $\gamma^{\text{weak}}_{L_k}$ down-weights weak experience, for $k \in 1, 2$.

C. Candidate Generation

Given $x_i = (h_i, D_i, P_i, O_i)$, HEM-Med uses $\mathcal{A}_i = D_i \cup P_i \cup O_i$ to retrieve medication–context experience from L_1 cross-hospital and L_2 hospital-specific residual memories, forming a candidate medication pool.

For each medication m, we compute the memory-based experience score as:

$$s_{\text{mem}}(m \mid x_i) = \alpha_1 g_{L_1}(m, \mathcal{A}_i) + \alpha_2 g_{L_2}(m, \mathcal{A}_i, h_i), \tag{19}$$

where g_{L_1} and g_{L_2} denote the aggregated experience scores from L_1 and L_2 memory, respectively.

To reduce the bias toward globally frequent medications, HEM-Med further incorporates prior statistics:

$$\Phi_{\text{prior}}(m, h_i, \mathcal{A}_i) = \phi_g \Pr(m) + \phi_h \Pr_{h_i}(m) + \phi_a \Pr(m \mid \mathcal{A}_i) + \phi_{ha} \Pr_{h_i}(m \mid \mathcal{A}_i). \tag{20}$$

where $\Phi_{\text{prior}}(m, h_i, \mathcal{A}_i)$ scores medication m using global, hospital, anchor, and hospital–anchor frequencies.

The initial candidate score is defined as:

$$s_{\text{init}}(m \mid x_i) = s_{\text{mem}}(m \mid x_i) + \beta \Phi_{\text{prior}}(m, h_i, \mathcal{A}_i) + b_{\text{comp}}(m) - \lambda_{\text{cand}} \psi_{\text{cand}}(m), \tag{21}$$

where $b_{\text{comp}}(m)$ compensates for under-retrieval, and $\psi_{\text{cand}}(m)$ denotes a DDI-based soft safety penalty.

Top-ranked medications form the candidate pool:

$$B_i = \text{Top}_B\left(\{s_{\text{init}}(m \mid x_i) \mid m \in \mathcal{M}\}\right), \tag{22}$$

where B is the retained candidate size. The candidate pool B_i is then passed to the learned experience reranker for further calibration and selection.

D. Experience Reranking and Safety Review

HEM-Med estimates $p_{\text{final}}(m \mid x_i)$ for $m \in B_i$ from calibrated reranking and predicts $p_{\text{size}}(k \mid x_i)$. Safety-constrained DDI-aware set decoding mitigates unsafe or redundant combinations from independent Top-K selection. Here, $\rho(u, v)$ and $\text{sev}(u, v)$ denote empirical co-administration confidence and DDI severity, respectively; $\text{Red}(Y)$ counts redundant medications within therapeutic categories.

HEM-Med generates the final medication set by optimizing the following set-level objective:

$$S(Y \mid x_i) = \sum_{m \in Y} \log p_{\text{final}}(m \mid x_i) + \log p_{\text{size}}(|Y| \mid x_i) - \lambda_{\text{DDI}} \sum_{\substack{u,v \in Y \\ u<v}} \text{sev}(u,v)(1-\rho(u,v)) - \lambda_{\text{red}} \text{Red}(Y). \tag{23}$$

The recommended medication set is obtained as:

$$\hat{Y}_i = \arg\max_{Y \subseteq B_i} S(Y \mid x_i). \tag{24}$$

In this objective, the first term encourages clinically relevant medications, and the second term favors a reasonable medication set size. The third term penalizes potentially unsafe DDI pairs, where the penalty increases with higher interaction severity and lower empirical co-administration confidence. The last term discourages redundant medications from the same therapeutic category.

E. Model Training

HEM-Med adopts a modular training strategy for candidate calibration and set-size estimation. The pointwise scorer, pairwise reranker, therapeutic-class scorer, and medication-count predictor are trained as separate submodules and are combined only during inference. Given the candidate set B_i and the ground-truth medication set M_i, each candidate medication $m \in B_i$ is assigned a binary label:

$$y_{i,m} = \mathbb{I}(m \in M_i). \tag{25}$$

We optimize medication-level prediction with a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{pt}} = \sum_i \sum_{m \in B_i} w_{i,m} \text{BCE}\left(y_{i,m}, p_{\text{pt}}(m \mid x_i)\right), \tag{26}$$

where $w_{i,m}$ alleviates medication-frequency imbalance.

To improve candidate ranking, we further construct positive-negative medication pairs and apply a pairwise ranking loss:

$$\mathcal{L}_{\text{pw}} = \sum_i \sum_{(m^+,m^-) \in \Pi_i} \text{softplus}\Big(- \big[r_{\text{pw}}(m^+ \mid x_i) - r_{\text{pw}}(m^- \mid x_i) \big] \Big) \tag{27}$$

where $\Pi_i = \{(m^+, m^-) \mid m^+ \in M_i \cap B_i,\ m^- \in B_i \setminus M_i\}$.

Since HEM-Med predicts the medication set size rather than using a Top-K threshold, the set-size predictor is trained by:

$$\mathcal{L}_{\text{size}} = -\sum_i \log p_{\text{size}}(|M_i| \mid x_i). \tag{28}$$

Pointwise, pairwise, therapeutic-class, and medication-count outputs are combined only at inference for calibrated reranking and DDI-aware set decoding. DDI knowledge serves only candidate-level safety penalties and final decoding, not additional supervision.

V. Experiment

A. Dataset

We evaluate HEM-Med on the eICU Collaborative Research Database, a real-world multi-center ICU database with over 200,000 de-identified ICU admissions in the United States [25]. In the processed cohort, the ground-truth medication sets have a DDI Rate of approximately 11.66%. Dataset statistics are reported in Table I.

TABLE I: Statistics of the processed eICU dataset.

Item	Number
# of records / hospitals	90,545 / 149
Avg. / max / min # of records per hospital	607.68 / 3,315 / 1
Diag. / proc. / med. vocabulary size	2,958 / 1,852 / 161
Lab vocabulary size	12
Avg. / max / min # of diagnoses per record	4.04 / 91 / 1
Avg. / max / min # of procedures per record	6.60 / 102 / 1
Avg. / max / min # of medications per record	14.36 / 63 / 0
Avg. / max / min # of lab features per record	3.76 / 12 / 0
# of records in train / validation / test	72,436 / 9,054 / 9,055

B. Evaluation Metrics

Following prior studies [1], [2], [13], we assess accuracy using Jaccard and F1 for set overlap and PRAUC for prediction ranking, and safety using DDI Rate, the proportion of potentially interacting predicted medication pairs. To report the overall performance, we introduce SafeScore:

$$\text{SafeScore} = \frac{\text{PRAUC} + \text{F1} + \text{Jaccard}}{3} \cdot (1 - \text{DDI}). \tag{29}$$

Higher SafeScore indicates a better accuracy-safety trade-off.

C. Implementation Details

We split the dataset into training, validation, and test sets with a ratio of 8:1:1. To avoid information leakage, all memory statistics and prior distributions are constructed only from the training set. For $\mathcal{K}^{(L_1)}$, the minimum anchor support threshold τ_a, anchor–medication support threshold τ_{am}, conditional frequency threshold τ_p, anchor–medication hospital-coverage threshold τ_h, and anchor hospital-coverage threshold τ_{ha} are set to 30, 10, 0.02, 4, and 5, respectively. For $\mathcal{K}^{(L_2)}$, the local anchor support threshold $\tau_a^{(L_2)}$, local anchor–medication support threshold $\tau_{am}^{(L_2)}$, local conditional frequency threshold $\tau_p^{(L_2)}$, and residual magnitude threshold $\tau_\Delta^{(L_2)}$ are set to 8, 4, 0.10, and 0.12, respectively. These thresholds are used as reliability filters for memory construction rather than primary model-selection hyperparameters.

During candidate generation, the candidate depth B is set to 60, and the weak-experience decay factors are set to $\gamma_{\text{weak}}^{(L_1)} = \gamma_{\text{weak}}^{(L_2)} = 0.42$. Since B and the weak-experience decay factors directly affect candidate recall and noise, we analyze them in the parameter study. All other parameters are selected on the validation set and kept fixed during testing.

For fair comparison, all baselines use the same data split, medication vocabulary, DDI file [8], [9], and random seed. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU with 24GB of memory.

TABLE II: Main experimental results on the eICU dataset.

Model	PRAUC ↑	F1 ↑	Jaccard ↑	DDI Rate ↓	SafeScore ↑
EMCDR	0.3421 ± 0.0043	0.3716 ± 0.0086	0.2647 ± 0.0095	0.1321 ± 0.0062	0.2831 ± 0.0044
DDTCDR	0.3598 ± 0.0136	0.3763 ± 0.0019	0.2665 ± 0.0102	0.1231 ± 0.0049	0.2931 ± 0.0053
TEMPT	0.3430 ± 0.0065	0.3335 ± 0.0054	0.2284 ± 0.0075	0.1204 ± 0.0038	0.2653 ± 0.0035
Reason4Rec	0.3697 ± 0.0083	0.3893 ± 0.0022	0.2783 ± 0.0036	$\mathbf{0.0814 \pm 0.0064}$	0.3176 ± 0.0036
EXP3RT	0.3732 ± 0.0148	0.3944 ± 0.0026	0.2819 ± 0.0028	0.0927 ± 0.0014	0.3174 ± 0.0046
SafeDrug	0.3926 ± 0.0038	0.4082 ± 0.0061	0.2939 ± 0.0049	0.0862 ± 0.0064	0.3334 ± 0.0035
VITA	0.3578 ± 0.0037	0.3782 ± 0.0025	0.2688 ± 0.0120	0.1107 ± 0.0071	0.2979 ± 0.0045
RES-MR	0.3697 ± 0.0046	0.3893 ± 0.0019	0.2782 ± 0.0076	0.1013 ± 0.0021	0.3107 ± 0.0028
HEM-Med	$\mathbf{0.4161 \pm 0.0086}$	$\mathbf{0.4532 \pm 0.0062}$	$\mathbf{0.3246 \pm 0.0074}$	0.0831 ± 0.0035	$\mathbf{0.3649 \pm 0.0042}$

The best results are marked in bold and the second-best results are underlined.

TABLE III: Ablation study results on the eICU dataset.

Variant	PRAUC ↑	F1 ↑	Jaccard ↑	DDI Rate ↓	SafeScore ↑
HEM-Med w/o MD	0.3723 ± 0.0145	0.4221 ± 0.0019	0.2938 ± 0.0092	0.0993 ± 0.0187	0.3267 ± 0.0085
HEM-Med w/o DDI-Dec	0.3631 ± 0.0136	0.4175 ± 0.0021	0.2840 ± 0.0102	0.1254 ± 0.0074	0.3104 ± 0.0056
HEM-Med w/o RERANKER	0.4025 ± 0.0049	0.3235 ± 0.0078	0.2111 ± 0.0053	0.0953 ± 0.0049	0.2826 ± 0.0036
HEM-Med w/o WAM	0.3677 ± 0.0047	0.4117 ± 0.0091	0.2824 ± 0.0062	0.0886 ± 0.0037	0.3226 ± 0.0039
HEM-Med	$\mathbf{0.4161 \pm 0.0086}$	$\mathbf{0.4532 \pm 0.0062}$	$\mathbf{0.3246 \pm 0.0074}$	$\mathbf{0.0831 \pm 0.0035}$	$\mathbf{0.3649 \pm 0.0042}$

D. Result Analysis

Table II reports results on the preprocessed eICU dataset. Baselines include transfer learning methods EMCDR and DDTCDR [5], [6], general LLM recommenders Reason4Rec and EXP3RT [26], [27], and MR methods SafeDrug, VITA, TEMPT, and RES-MR [2], [7], [28], [29]. TEMPT targets multi-center MR. All methods undergo 10 bootstrap evaluations with a sampling ratio of 0.8; metric means and standard deviations are reported.

HEM-Med outperforms transfer and general LLM baselines on all accuracy-related metrics and achieves the highest SafeScore. These results suggest that domain alignment or preference reasoning alone may be insufficient for multi-center MR. Compared with SafeDrug, the strongest MR baseline, HEM-Med improves the accuracy–safety trade-off through disentangled memory, calibrated reranking, and DDI-aware decoding. Although Reason4Rec has the lowest DDI Rate, HEM-Med combines higher accuracy with competitive safety.

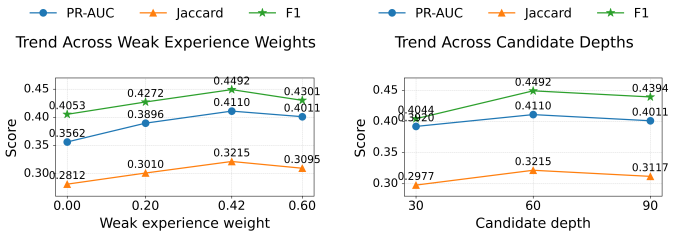
E. Ablation Study

Table III reports four ablations: w/o MD, w/o DDI-Dec, w/o RERANKER, and w/o WAM, removing memory decoupling, DDI-aware decoding, experience reranking, and weak auxiliary memory, respectively. Each removal degrades performance, supporting the contributions of memory disentanglement, safety-aware decoding, candidate calibration, and weak experience. Full HEM-Med achieves the best overall performance.

F. Parameter Analysis

We analyze the influence of two key hyperparameters: the weak auxiliary experience weight and the candidate depth. The results are shown in Fig. 3.

Impact of weak auxiliary experience weight. With the candidate depth fixed to 60, introducing weak auxiliary experience improves recommendation performance compared with removing weak memory. The model achieves the best performance when the weak experience weight is set to 0.42. Further increasing the weight leads to performance degradation, indicating that weak experience is useful as supplementary information but may introduce noise when overemphasized.



(a) Weak weight (b) Candidate depth

Fig. 3: Parameter analysis of weak experience weight and candidate depth.

Impact of candidate depth. With the weak experience weight fixed to 0.42, the model performs best when the candidate depth is set to 60. A smaller candidate depth may miss relevant medications before reranking, while a larger depth introduces more noisy candidates and false positives. Therefore, we set the weak auxiliary experience weight to 0.42 and the candidate depth to 60 in the final experiments.

G. Case Study

Table IV presents two ICU cases from different hospitals. Both cases involve alcohol or toxicology-related conditions and electrolyte management, yet their medication requirements differ. Although the baseline methods identify some common electrolyte medications, they frequently produce false-positive recommendations and miss key medications, such as *Furosemide* in Case A and *Hydralazine* in Case B, indicating limited adaptation to hospital- and patient-specific needs. In contrast, HEM-Med accurately matches the medication sets in both cases, preserving shared electrolyte medications

TABLE IV: Case study comparison between two clinically similar ICU stays from different hospitals.

Method	Case A: H452	Case B: H444
Ground Truth	Dex; Fur; NaCl; KCl; $MgSO_4$	Enox; Hyd; Lor; NaCl; KCl
DDTCDR	**3/5** NaCl; $MgSO_4$; CaGlu; KCl; APAP *Miss:* Dex; Fur	**3/5** KCl; Ond; APAP; NaCl; Lor *Miss:* Enox; Hyd
SafeDrug	**4/5** NaCl; CaGlu; $MgSO_4$; KCl; Dex *Miss:* Fur	**4/5** KCl; Lor; Eso; NaCl; Enox *Miss:* Hyd
TEMPT	**3/5** NaCl; Lor; Dex; KCl; Glu *Miss:* Fur; $MgSO_4$	**4/5** Enox; Lor; NaCl; KCl; Eso *Miss:* Hyd
RES-MR	**4/5** NaCl; KCl; $MgSO_4$; CaGlu; Dex *Miss:* Fur	**4/5** KCl; Lor; NaCl; Eso; Enox *Miss:* Hyd
HEM-Med	**5/5** **NaCl; KCl; $MgSO_4$; Dex; Fur** *Miss:* None	**5/5** **Lor; Enox; KCl; Hyd; NaCl** *Miss:* None

Color legend: Blue = correctly predicted medication; Red = false-positive medication; Orange = missed ground-truth medication.
Abbreviations: Dex = Dextrose; Fur = Furosemide; NaCl = Sodium chloride; KCl = Potassium chloride; $MgSO_4$ = Magnesium sulfate; CaGlu = Calcium gluconate; APAP = Acetaminophen; Ond = Ondansetron; Eso = Esomeprazole; Glu = Glucagon; Enox = Enoxaparin; Hyd = Hydralazine; Lor = Lorazepam.

while identifying case-specific requirements. This demonstrates its ability to integrate cross-hospital common knowledge, hospital-specific experience, and safety-aware decoding.

VI. Conclusion

In this paper, we propose HEM-Med, a hierarchical agentic memory framework for safe multi-center medication recommendation. HEM-Med explicitly separates cross-hospital consensus experience from hospital-specific residual patterns and integrates DDI-aware decoding for safer medication set generation. Experiments on a real-world multi-center eICU dataset show that HEM-Med improves recommendation accuracy while maintaining medication safety. Ablation studies validate the effectiveness of its key components. Future work will explore richer clinical constraints, fine-grained medication safety knowledge, and prospective validation.

Acknowledgment

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), the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No. 2025ZDZX01).

References

[1] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, "Gamenet: Graph augmented memory networks for recommending medication combination," in *AAAI*, vol. 33, 2019, pp. 1126–1133.

[2] C. Yang, C. Xiao, F. Ma, L. Glass, and J. Sun, "Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations," in *IJCAI*, 2021, pp. 3735–3741.

[3] Y. Zhao, P. Wang, and X. Shi, "Dkgm: A diffusion knowledge graph model for medication recommendation leveraging medical and drug information," in *ICHI*, 2025, pp. 268–279.

[4] C. Wang, S. Wang, Y. Li, Q. Zhang, L. Hu, and W. Lu, "Eval4direc: A unified and systematic evaluation framework for diffusion-based recommender systems," in *ACM TKDD*, 2026, p. 3856821.

[5] T. Man, H. Shen, X. Jin, and X. Cheng, "Cross-domain recommendation: An embedding and mapping approach." in *IJCAI*, vol. 17, 2017, pp. 2464–2470.

[6] P. Li and A. Tuzhilin, "Ddtcdr: Deep dual transfer cross domain recommendation," in *WSDM*, 2020, pp. 331–339.

[7] Q. Liu, Z. Qiu, X. Zhao *et al.*, "A contrastive pretrain model with prompt tuning for multi-center medication recommendation," *ACM Transactions on Information Systems*, vol. 43, no. 3, pp. 1–29, 2025.

[8] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science Translational Medicine*, vol. 4, no. 125, pp. 125ra31–125ra31, 2012.

[9] D. S. Wishart, Y. D. Feunang, A. C. Guo *et al.*, "Drugbank 5.0: A major update to the drugbank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[10] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NeurIPS*, vol. 29, 2016, p. 1.

[11] H. Le, T. Tran, and S. Venkatesh, "Dual control memory augmented neural networks for treatment recommendations," in *PAKDD*, 2018, pp. 273–284.

[12] Y. Li, Q. Zhang, W. Lu *et al.*, "Time-aware medication recommendation via intervention of dynamic treatment regimes," in *ACM Web Conference*, 2025, pp. 5161–5172.

[13] N. Yang, K. Zeng, Q. Wu, and J. Yan, "Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning," in *ACM Web Conference*, 2023, pp. 4075–4085.

[14] S. Liu, X. Wang, X. Zhao, and H. Chen, "Dkinet: Medication recommendation via domain knowledge informed deep learning," in *BIBM*, 2024, pp. 3521–3526.

[15] X. Dong, Y. Zheng, Z. Shu *et al.*, "Tcmpr: Tcm prescription recommendation based on subnetwork term mapping and deep learning," *BioMed Research International*, vol. 2022, no. 1, p. 4845726, 2022.

[16] H. Yang, Y. Zhang, J. Zhang, K. Che, and L. Xing, "Secgi: A medication recommendation model combining llm-based semantic enrichment and causal graph inference," in *BIBM*, 2025, pp. 3130–3136.

[17] J. Cao, X. Lin, X. Cong, J. Ya, T. Liu, and B. Wang, "Disencdr: Learning disentangled representations for cross-domain recommendation," in *ACM SIGIR*, 2022, pp. 267–277.

[18] K. Singhal, S. Azizi, T. Tu *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[19] X. Tang, A. Zou, Z. Zhang *et al.*, "Medagents: Large language models as collaborators for zero-shot medical reasoning," in *ACL*, 2024, pp. 599–621.

[20] K. Chen, X. Li, T. Yang, H. Wang, W. Dong, and Y. Gao, "Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation," *arXiv preprint arXiv:2503.13856*, 2025.

[21] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li, "Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation," *arXiv preprint arXiv:2306.09968*, 2023.

[22] S. Gao, R. Zhu, Z. Kong *et al.*, "Txagent: An ai agent for therapeutic reasoning across a universe of tools," *arXiv preprint arXiv:2503.10970*, 2025.

[23] K. Liu, Z. Li *et al.*, "Jointcoder: Exploring automated icd coding on real-world chinese ehrs with a multi-agent framework," in *ACL*, 2026, pp. 263–274.

[24] E. Wang, J. Wang *et al.*, "Soaptriage: Soap-guided multi-view clinical text modeling framework for automated esi prediction," in *ACL*, 2026, pp. 34 813–34 837.

[25] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, no. 1, p. 180178, 2018.

[26] Y. Fang, W. Wang, Y. Zhang *et al.*, "Reason4rec: Large language models for recommendation with deliberative user preference alignment," *arXiv preprint arXiv:2502.02061*, 2025.

[27] J. Kim, H. Kim, H. Cho *et al.*, "Review-driven personalized preference reasoning with large language models for recommendation," in *ACM SIGIR*, 2025, pp. 1697–1706.

[28] T. Kim, J. Heo, H. Kim, K. Shin, and S.-W. Kim, "Vita:'carefully chosen and weighted less' is better in medication recommendation," in *AAAI*, vol. 38, no. 8, 2024, pp. 8600–8607.

[29] C. Wang, J. Li, S. Wang, Y. Li, H. Gu, and W. Lu, "Res-mr: Risk-aware reasoning for explainable and safe medication recommendation," in *ACM SIGIR*, 2026, accepted paper.