TRIXMED: TRIAGE-ROUTED MIXTURE OF EXPERTS FRAMEWORK FOR INTERPRETABLE DRUG RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Drug recommendation is a critical task in intelligent healthcare systems that significantly impacts patient outcomes. While large language models (LLMs) have advanced the field through sophisticated semantic understanding, current approaches face two fundamental challenges: (1) they fail to adequately address patient heterogeneity, treating diverse populations with a one-size-fits-all model; (2) their black-box nature undermines clinical trust and adoption. We introduce TRIXMED (Triage-Routed Interpretable eXpert Medicine), a novel framework that integrates Mixture of Experts (MoE) architecture with routing mechanisms that mimic clinical triage processes for personalized drug recommendation. TRIXMED addresses patient heterogeneity by introducing specialized experts that handle distinct patient subgroups, while ensuring interpretability through a clustering-based routing strategy that automatically directs patients to the most appropriate expert based on their clinical profile. Our approach employs a unique warm-up training phase followed by feature extraction and patient stratification, enabling transparent expert routing based on patient characteristics. Extensive experiments on the MIMIC-III datasets show that TRIXMED surpasses the SOTA model, achieving relative improvements of 27.4% in Jaccard index, and 15.61% in F1-score, respectively. TRIXMED represents a significant advancement in bridging the gap between AI-powered recommendations and clinical practice through its combination of heterogeneity handling and transparent decision-making.

1 INTRODUCTION

The proliferation of electronic health records (EHRs) (Yadav et al., 2018) has created unprecedented opportunities for developing predictive systems to enhance clinical decision-making. Medication recommendation represents a particularly critical application, enabling physicians to optimize drug regimens for patients with complex comorbidities (World Health Organization, 2023; U.S. Food and Drug Administration, 2023). The evolution of medication recommendation exemplifies a trajectory from classic collaborative filtering (Su & Khoshgoftaar, 2009; Morales et al., 2022) to modern deep learning paradigms, notably Recurrent Neural Networks (RNNs) (Medsker et al., 2001) and Transformers (Vaswani et al., 2017), driven by the goal of effectively modeling patient temporal dynamics. Recently, Large Language Models(LLMs) based methods have demonstrated superior pattern recognition capabilities (Liu et al., 2024; Hassan & Elagamy, 2025; Lu et al., 2025), they introduce critical challenges:

First, current mainstream medication recommendation models typically rely on population-level modeling and fail to effectively distinguish between distinct clinical phenotypes, leading to suboptimal performance when dealing with patients with rare diseases or complex comorbidities. For patient groups suffering from multiple chronic conditions or presenting atypical clinical symptoms, treatment plans must fully account for their unique pathophysiological characteristics. If a system lacks the ability to identify subtle patient subgroups, it may generate recommendation misaligned with individual pathological mechanisms, potentially compromising treatment safety and efficacy. Second, the decision-making process of these models lacks transparency and does not provide clinically meaningful explanatory mechanisms, making it difficult for users to trace the underlying logic behind specific medication recommendations. In actual clinical practice, healthcare professionals require not only accurate recommendations but also verifiable decision support, such as trends in key clinical indicators, patterns of historical treatment responses, or correlations related to drug mechanisms. The lack of transparency in "black-box" systems undermines clinical trust and willingness to adopt, particularly in high-risk medical scenarios where medication errors could lead to serious consequences, justifying a cautious approach.

To address these challenges, we present TRIXMED (Triage-Routed Interpretable eXpert Medicine), a novel framework that integrates a Mixture-of-Experts (MoE) (Shazeer et al., 2017) architecture with triage-inspired clinical routing. TRIXMED begins by transforming structured medical codes from patient visit histories into natural language descriptions, enabling contextual understanding and semantic reasoning while improving generalization to previously unseen patients. The model construction follows a three-phase approach. In the phase 1, TRIXMED trains a warm-up model using 30% of the training data to learn initial patient representations from both the naturalized clinical narratives and structured EHRs features. This warm-up model then extracts comprehensive features from the entire training dataset. In the phase 2, these extracted features enable unsupervised clustering

to identify clinically coherent patient subgroups based on their medical profiles. In the phase 3, the clustering results inform the MoE architecture, establishing a routing mechanism that dynamically assigns patient records to specialized expert modules according to their subgroup membership. Each expert module specializes in the treatment patterns and risk factors specific to its assigned patient cluster. This design not only improves the personalization and effectiveness of drug recommendations, but also enhances interpretability by explicitly aligning expert selection with clinically transparent patient characteristics. Our contributions are outlined as follows:

(1) We propose TRIXMED, the first medication recommendation system that automatically identifies and adapts to distinct patient subgroups through clinical feature clustering. Unlike existing monolithic approaches, our framework employs specialized experts trained on clustered EHRs data partitions, enabling personalized therapeutic recommendations for diverse clinical phenotypes.

(2) We design a transparent patient-to-expert assignment strategy that mimics clinical triage processes. By combining warm-up feature extraction with unsupervised clustering, our routing mechanism creates clinically meaningful decision boundaries based on inherent patient characteristics.

(3) We establish a comprehensive evaluation system to validate the effectiveness of TRIXMED. Our approach demonstrates improvements of 27.4% and 15.61% in Jaccard and F1 scores, respectively, on the MIMIC-III datasets (Johnson et al., 2016) .

2 Related Work

Graph-Based Drug Recommendation Early works in medication recommendation focused on modeling medical entity relationships through graph architectures. TAHDNet (Su et al., 2022) integrates Transformer modules , 1D-CNN, and dynamic time-aware mechanisms to capture disease progression patterns. Subsequent works introduced explicit drug interaction modeling: REFINE (Bhoi et al., 2023) is a severity-weighted interaction graphs with balanced loss functions but neglect unstructured clinical data integration. ACDNet (Mi et al., 2024) constructs an innovative attention-guided collaborative decision-making framework, merging patient electronic health records with drug molecular graph features. However, its multi-module structure increases computational complexity and limits the effectiveness of safety optimization. StratMed (Li et al., 2024) proposes a hierarchical reinforcement strategy to balance drug safety and efficacy accuracy, achieving efficient training under the long-tail distribution of medical data through a dual-attribute graph network, but it lacks adaptability to extremely sparse data. COGNet (Wu et al., 2022) enhances drug recommendation through a novel copy-predict mechanism within an encoder-decoder framework. The encoder assimilates both current and historical patient diagnostic data, while the decoder, integrated with a Graph Convolutional Network (GCN), models drug co-occurrence and drug-drug interactions (DDI). At the visit level, it filters similar historical visits, and at the drug level, it employs attention mechanisms to select reusable drugs, dynamically determining whether to replicate historical drugs or generate new ones.

Generative Drug Recommendation Generative approaches employ advanced architectures for complex decision patterns. The 4SDrug framework (Tan et al., 2022) adopts a set-based generative approach, formulating drug recommendation as a set-to-set combinatorial optimization problem. By leveraging inter-set similarity and integrating knowledge-driven and data-driven penalties, it generates optimized drug combinations that prioritize both safety and clinical relevance. PharmaBERT (Jayapradha et al., 2024) showcases the potential of pre-trained models in drug sentiment analysis, with visualization tools enhancing decision transparency, yet its computational demands and adaptation to specialized terminology highlight current technological limitations. Generative techniques offer a new paradigm for drug recommendation. LEADER (Liu et al., 2024) introduces a feature-level knowledge distillation framework, transferring the semantic understanding of LLMs to lightweight models, effectively addressing the cold start problem, though it remains constrained by pre-training quality and the risk of information loss. The ExpDrug model (Lu et al., 2025) employs a feature mapping method based on interpretable dimensional projection to map implicit patient state vectors to clinically meaningful decision dimensions, enabling interpretable medication recommendation. A controllable threshold strategy is also proposed, which adds a dynamic DDI suppression term to the loss function to explicitly reduce recommendations of high-interaction-risk drug combinations during optimization.

3 Methodology

As illustrated in Figure 1, TRIXMED model comprises of four core components: (1) A Data Mapping and Integration module that employs a mapping framework to convert EHRs into coherent natural language descriptions; (2) A Patient-Specific Feature Extraction module, that pre-trains a warm-up model on 30% of the training data and subsequently uses it to derive comprehensive feature representations from the full dataset; (3) A Patient Stratification module that groups the extracted features into clinically meaningful patient subgroups based on their learned representations; (4) A triage-based expert routing module that dynamically assigns each patient to the most appropriate specialized expert network according to clustering outcomes, ensuring that cases with similar

clinical profiles are handled by consistent experts for personalized treatment recommendations. Diverging from conventional MoE, TRIXMED introduces a clustering-guided routing strategy that aligns patient subgroups with dedicated therapeutic experts, emulating a clinical triage process to enhance specialization and personalization.

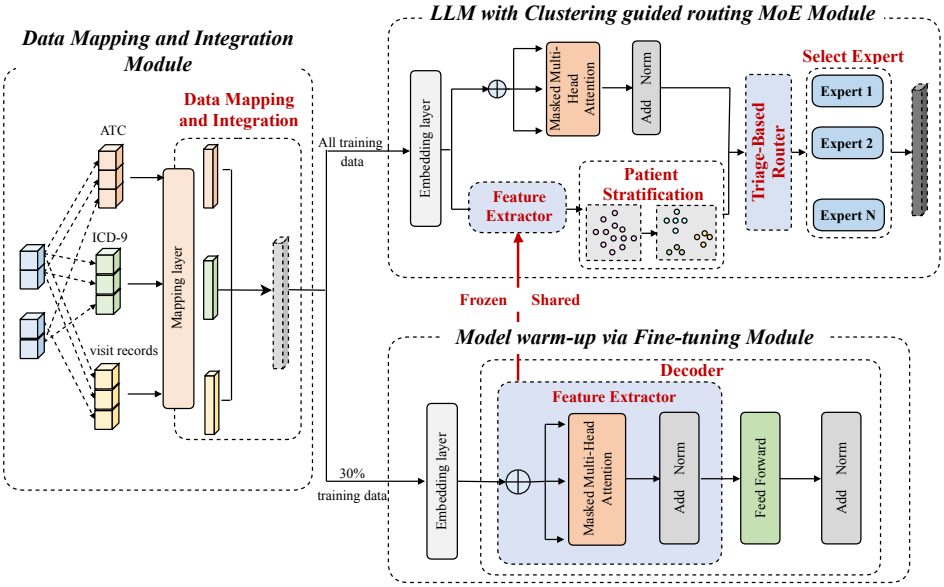


Figure 1: Framework of TRIXMED

Problem Statement Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ containing longitudinal EHRs from N patients. Medical history of each patient comprises a variable-length sequence $\mathcal{H}_p = \langle e^{(1)}, \ldots, e^{(T-1)} \rangle$, where each encounter $e^{(i)} = (d^{(i)}, p^{(i)}, m^{(i)})$ encapsulate the triplet of diagnoses, procedures, and medications recorded at time i. At the current time T, the visit record $e^{(T)} = (d^{(T)}, p^{(T)})$ contains only diagnostic and procedural information, while the corresponding medication set $m^{(T)}$ constitutes the prediction target y_i. The goal is to learn a recommendation function $f(\cdot)$ that given historical visits $\mathbf{x}_i = [\mathcal{H}_p, d^{(T)}, p^{(T)}]$, it predicts: $\hat{m}^{(T)} = f(\mathbf{x}_i) \in \{0, 1\}$, where the function maps patient history and current diagnoses/procedures to prescribed medications.

To achieve this objective while addressing patient heterogeneity, we employ a three-stage fine-tuning strategy:

(1) Warm-up Feature Learning: A preliminary encoder f_{warm} is trained on 30% of dataset $\mathcal{D}$ to capture fundamental patient patterns. Subsequently, this trained model extracts comprehensive feature representations $\{\mathbf{z}_i\}_{i=1}^{N}$ from all training instances, where $\mathbf{z}_i = f_{warm}(\mathbf{x}_i)$ encodes clinical profile of patient i.

(2) Patient Stratification: We apply Principal Component Analysis (PCA) (Abdi & Williams, 2010) for dimensionality reduction followed by K-means (Ahmed et al., 2020) clustering on $\{\mathbf{z}_i\}_{i=1}^{N}$ to identify K patient subgroups: $\mathcal{G} = \{G_1, \ldots, G_K\}$, where $G_k \subset \mathcal{D}$ represents the k-th patient cluster.

(3) Expert Triage Routing: A MoE network $\mathcal{E} = \{E_k\}_{k=1}^{K}$ is constructed, where each expert E_k specializes in the therapeutic patterns of patient cluster G_k. Instead of assigning each input to a single expert, we select the top-n most relevant experts, where n indicates the number of activated experts. The routing function generates a sparse weight vector where only the entries corresponding to the top-n experts are non-zero.

3.1 Data Mapping and Integration

Longitudinal EHRs data are processed through a structured pipeline to reconstruct patient health trajectories. This pipeline comprises two integral stages: code-to-language transformation and temporal sequence construction, which together facilitate the semantic alignment and temporal modeling of clinical events.

Code-to-Language Transformation Raw medical codes are devoid of the semantic context required for accurate comprehension by LLMs. Consequently, we construct bidirectional mappings between structured medical codes and natural language descriptions, which facilitates contextual reasoning and the inference of inter-entity relationships.

(1) Medication Mapping. The Anatomical Therapeutic Chemical (ATC) classification system (Chen et al., 2012) provides hierarchical drug coding. We formalize the mapping as:

$$f_{\text{ATC}} : \mathcal{C}_{\text{med}} \rightarrow \mathcal{M},$$

where $\mathcal{C}_{\text{med}}$ denotes ATC codes and $\mathcal{M}$ represents natural language medication names.

(2) Diagnosis / Procedure Mapping. For diagnoses and procedures, we leverage the ICD-9/10 codes (Quan et al., 2005) and define a bidirectional mapping to standardized clinical descriptions:

$$f_{\text{ICD}} : \mathcal{C}_{\text{diag/proc}} \rightarrow \{\mathcal{D}, \mathcal{P}\},$$

where $\mathcal{C}_{\text{diag/proc}}$ contains ICD codes, while $\mathcal{D}$ and $\mathcal{P}$ denote the natural language descriptions of diagnoses and procedures, respectively.

Temporal Sequence Construction Each historical visit $e^{(t)} = \left(d^{(t)}, p^{(t)}, m^{(t)}\right)$, including diagnosis descriptions $d^{(t)} = f_{\text{ICD}}(C^{(t)}_{\text{diag}})$, procedure descriptions $p^{(t)} = f_{\text{ICD}}(C^{(t)}_{\text{proc}})$, and medication names $m^{(t)} = f_{\text{ATC}}(C^{(t)}_{\text{med}})$. The task is to predict the medication set $m^{(T)}$ based on the historical sequence and the current diagnosis and procedure.

3.2 Patient Feature Extraction

Conventional methods often ignore patient heterogeneity due to their inability to capture diverse patient subgroups, resulting in suboptimal recommendations. To address this, we learn personalized patient representations by training a transformer-based warm-up model on a subset of data. This model captures semantic and temporal patterns from medical sequences. Features are then extracted for the entire cohort, aggregated, and reduced via PCA to produce informative low-dimensional embeddings suitable for clustering and stratification.

Warm-up A warm-up model is trained on a subset of the data ($\mathcal{D}_{\text{warm}} \subset \mathcal{D}$, 30%), to learn foundational patient representations that capture clinically meaningful patterns. For each patient instance $(x_i, y_i) \in \mathcal{D}_{\text{warm}}$, the raw medical codes are first transformed into semantic embeddings through function f_{embed}:

$$e_i = f_{\text{embed}}(x_i) = \left[\tilde{\mathbf{e}}^{(t)}\right]_{t=1}^{T}, \quad \tilde{\mathbf{e}}^{(t)} = \begin{cases} [\mathbf{E}_d \cdot d^{(t)},\ \mathbf{E}_p \cdot p^{(t)},\ \mathbf{E}_m \cdot m^{(t)}], & \text{if } t < T \\ [\mathbf{E}_d \cdot d^{(t)},\ \mathbf{E}_p \cdot p^{(t)},\ \mathbf{0}], & \text{if } t = T \end{cases}$$

where $\mathbf{E}_d$, $\mathbf{E}_p$, $\mathbf{E}_m$ are learnable embedding matrices for diagnoses, procedures, and medications respectively. These embeddings convert discrete medical codes into continuous semantic representations.

The warm-up model employs a standard Transformer encoder (Vaswani et al., 2017) and optimizes the autoregressive loss:

$$\mathcal{L}_{\text{warm-up}} = -\frac{1}{|\mathcal{D}_{\text{warm}}|} \sum_{(x,y)\in\mathcal{D}_{\text{warm}}} \sum_{t=1}^{T} \log P(y_t | y_{<t}, e_i; \theta)$$

This loss function measures the likelihood of predicting the next token y_t given all previous tokens $y_{<t}$ and the embedded input sequence, parameterized by θ. Minimizing this loss encourages the model to learn coherent and clinically plausible sequences of medical events.

Feature Extraction Having trained the warm-up model on $\mathcal{D}_{\text{warm}}$, we leverage it as a feature extractor to derive informative patient representations from the entire dataset $\mathcal{D}$. This step is to transfer the clinically meaningful patterns learned during warm-up to the full patient cohort, enabling semantically rich encoding of heterogeneous medical sequences.

For each patient $(x_i, y_i) \in \mathcal{D}$, we first apply the embedding function f_{embed} to convert discrete medical codes into semantic embeddings. These embeddings are then processed through the frozen warm-up model. As the input passes through the L layers of the Transformer, each layer ℓ produces a set of intermediate contextual representations. From the final layer, we extract the sequence of hidden states: $\mathbf{H}_i = [h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(T)}] \in \mathbb{R}^{T\times d}$, where $h_i^{(t)}$ represents the contextualized embedding for time step t and d is the hidden dimension. Since patient sequences vary in length, we aggregate temporal information into a fixed-size representation via mean pooling:

$$\bar{h}_i = \frac{1}{T}\sum_{t=1}^{T} h_i^{(t)} \in \mathbb{R}^d$$

This yields a d-dimensional vector capturing the complete clinical trajectory. However, clustering directly in the high-dimensional space $\mathbb{R}^d$ is computationally inefficient and prone to the curse of dimensionality, which can amplify noise and reduce clustering stability. To address this, we apply PCA to project each $\bar{h}_i$ into a lower-dimensional manifold:

$$z_i = \mathbf{W}_{\text{pca}} \cdot \bar{h}_i \in \mathbb{R}^k$$

where $\mathbf{W}_{\text{pca}} \in \mathbb{R}^{k\times d}$ contains the top-k principal components. This dimensionality reduction preserves the most informative variance while enabling efficient clustering. The final feature set $\{z_i\}_{i=1}^N$ serves as input to K-means clustering for patient stratification.

3.3 PATIENT STRATIFICATION

Based on the low-dimensional feature representations $\{z_i\}_{i=1}^N$ derived from the preceding representation learning stage, we proceed to partition the patient cohort into clinically coherent subgroups. This approach aims to identify latent patient phenotypes characterized by shared clinical profiles, while simultaneously structuring the training data for the MoE framework. By grouping patients into semantically meaningful clusters, the methodology facilitates targeted expert specialization and enhances ability of LLM to capture clinically relevant patterns.

Prior to clustering, we apply z-score normalization to ensure features are centered and scaled, mitigating the influence of heterogeneous feature magnitudes on distance-based clustering:

$$\tilde{z}_i = \frac{z_i - \mu_z}{\sigma_z + \epsilon}$$

where μ_z and σ_z denote the empirical mean and standard deviation computed across all patient features, with ϵ for numerical stability. We apply K-means clustering on the standardized features to identify K distinct patient strata:

$$\mathcal{G} = \text{K-means}(\{\tilde{z}_i\}_{i=1}^N, K)$$

resulting in a partition $\mathcal{G} = \{G_1, G_2, \ldots, G_K\}$, where each cluster G_k corresponds to a subgroup of patients exhibiting similar clinical characteristics.

Each patient i is assigned a cluster label $c_i \in \{1, \ldots, K\}$ based on minimal Euclidean distance to cluster centroids. Here, c_i denotes the index of the cluster $G_{c_i=k}$ to which patient i belongs. These assignments partition the training dataset $\mathcal{D}$ into cluster-specific subsets, enabling each expert model E_k to specialize in clinical patterns representative of its assigned subgroup.

3.4 TRIAGE-BASED EXPERT ROUTING

Having identified K distinct patient strata through clustering, we now introduce a triage-based routing mechanism that integrates these clinically derived groups into the MoE architecture. This design emulates real-world clinical decision-making, where patients are directed to specialized care units based on their specific conditions. By routing patients to experts according to learned phenotypic patterns, we aim to enhance both the specialization and interpretability of medication recommendations. The process consists of two core components: (1) cluster-guided expert activation, which maps patients to appropriate experts, and (2) adaptive prediction synthesis, which integrates specialized recommendations into a unified treatment plan.

Cluster-Guided Expert Activation The routing process begins with each patient's cluster assignment $c_i \in \{1, \ldots, K\}$ obtained from the stratification phase. a batch of B patients with assignment vector $\mathbf{c} \in \{1, \ldots, K\}^B$, we compute expert affinities through a learned linear transformation:

$$\mathbf{Z} = \text{OneHot}(\mathbf{c})\mathbf{W}_e + \mathbf{b}_e$$

where $\mathbf{W}_e \in \mathbb{R}^{K \times E}$ is a learnable weight matrix that maps cluster membership to expert relevance scores across E available experts. This transformation ensures that patients from the same clinical cluster exhibit similar expert activation patterns, maintaining consistency in therapeutic approach for phenotypically similar patients.

To balance specialized focus with computational efficiency, we implement a sparse routing mechanism through two sequential operations:

(1) Affinity Normalization. Convert raw relevance scores into probability distributions using row-wise softmax:

$$\mathbf{G} = \text{Softmax}(\mathbf{Z}) \in \mathbb{R}^{B \times E}$$

where $\mathbf{G}[i, e]$ represents the normalized relevance score of expert e for patient i.

(2) Top-k Selection. Select the k most relevant experts for each patient:

$$(\mathbf{\Gamma}, \mathbf{I}) = \text{TopK}(\mathbf{G}, k)$$

yielding expert weights $\mathbf{\Gamma}$ and indices $\mathbf{I}$. This sparse activation strategy ensures computational efficiency while preserving the flexibility to incorporate multiple expert perspectives when clinically warranted.

Each expert E_j processes only its assigned patients through selective computation:

$$\mathbf{H}_{\text{expert}}^{(j)} = E_j(\mathbf{H}_{\text{in}} \odot \mathbf{M}_j)$$

where $\mathbf{M}_j = \mathbb{I}(\mathbf{I} = j)$ is a binary mask that selects patients assigned to expert j (i.e., it equals 1 where $\mathbf{I} = j$ and 0 otherwise). This selective processing mirrors clinical triage systems where specialists focus on patients within their domain expertise. The learned cluster-to-expert mapping enables each expert to develop deep, specialized knowledge for its designated patient subgroup, significantly enhancing both prediction accuracy and clinical interpretability.

Adaptive Prediction Synthesis The final stage integrates specialized expert outputs into coherent medication recommendations. For each patient i, we synthesize recommendations from their top-k assigned experts using the learned routing weights:

$$\mathbf{H}_{\text{final}}[i] = \sum_{j=1}^{k} \mathbf{\Gamma}[i,j] \cdot \mathbf{H}_{\text{expert}}^{(\mathbf{I}[i,j])}[i]$$

where $\mathbf{I}[i,j]$ indexes the j-th selected expert for patient i, and $\mathbf{\Gamma}[i,j]$ represents the contribution weight of selected expert. Each expert-specific output $\mathbf{H}_{\text{expert}}^{(\mathbf{I}[i,j])}[i]$ encapsulates medication recommendations from a specialized clinical perspective.

When training the triage-based MoE architecture, a common issue is overly relies on a small subset of experts, thereby limiting model capacity and specialization (Shazeer et al., 2017). To mitigate the risk of routing collapse, we incorporate an auxiliary load $\mathcal{L}_{\text{aux}}$ balancing loss into the training objective.

$$\mathcal{L}_{\text{aux}} = E \sum_{e=1}^{E} f_e \cdot r_e, \quad f_e = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{expert } e \in \text{top-}n(\mathbf{G}[i])), \quad r_e = \frac{1}{N} \sum_{i=1}^{N} \mathbf{G}[i,e],$$

Here, f_e corresponds to the proportion of patients assigned to expert e through the top-n selection process, while r_e measures the average routing probability allocated to expert e across the batch. The final loss function combines the primary binary cross-entropy loss (Zhang & Sabuncu, 2018) with this balancing term:

$$\text{L}_{\text{final}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{BCE}} \left(m_i^{(T)}, \sigma(\mathbf{W}_o \mathbf{H}_{\text{final}}[i] + \mathbf{b}_o) \right) + \lambda \mathcal{L}_{\text{aux}}$$

This loss combines the primary prediction error with an expert balancing term. The binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ measures the discrepancy between the ground truth medications $m_i^{(T)}$ and predictions generated by transforming the synthesized patient representation $\mathbf{H}_{\text{final}}[i]$ through output weights $\mathbf{W}_o$, bias $\mathbf{b}_o$, and sigmoid activation σ. The auxiliary loss $\mathcal{L}_{\text{aux}}$, weighted by λ, promotes balanced expert utilization by encouraging uniform routing distributions across experts. This joint optimization ensures accurate medication predictions while maintaining effective expert specialization.

4 Experiments

4.1 Experimental Settings

We conduct comprehensive experiments on two widely-adopted EHRs datasets: MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2020). Both datasets contain detailed clinical records including patient diagnoses, procedures, and medication prescriptions. Following standard practice in medication recommendation literature, the datasets are partitioned into training $\mathcal{D}_{\text{train}}$ and testing $\mathcal{D}_{\text{test}}$ sets with a 7:3 ratio. The dataset includes both single-visit and multiple-visit patients, and the distribution proportion is detailed in the appendix.

We adopt a comprehensive set of metrics to evaluate both the accuracy and safety of medication recommendations: (1) Jaccard Similarity, measuring the overlap between predicted and ground-truth medication sets; (2) Precision, quantifying the fraction of recommended medications that are correct; (3) Recall, capturing the fraction of ground-truth medications successfully identified; and (4) F1 Score, the harmonic mean of precision and recall. For safety assessment, we evaluate (5) DDI Rate, the proportion of drug-drug interaction pairs among all possible medication combinations in the recommendation, where lower values indicate safer prescriptions; and (6) Average Number of Medications, monitoring the model's tendency towards polypharmacy.

Our model is based on the LLaMA3.1-8B (Touvron et al., 2023) foundation model, which is further fine-tuned and enhanced with a MoE architecture comprising 8 experts and a top-$n = 2$ sparse routing mechanism. We employ LoRA (Low-Rank Adaptation) (Hu et al., 2022) for efficient fine-tuning of the model, completing the training process with parameter-efficient adaptation. All experiments were conducted on an NVIDIA A800 GPU with 80GB memory, utilizing PyTorch 2.4.0 and CUDA 12.2.

4.2 Overall Performance

We comprehensively evaluate TRIXMED against nine baseline methods spanning different architectural paradigms: traditional deep learning models (RETAIN (Choi et al., 2016), LEAP (Zhang et al., 2017)), graph neural network approaches (GAMENet (Shang et al., 2019), COGNet (Wu et al., 2022), ACDNet (Mi et al., 2024), AMGNet (Li et al., 2025), (Tang et al., 2024)), safety-aware methods (SafeDrug (Yang et al., 2021), REFINE (Bhoi et al., 2023)), and LLM-based models (LEADER (Liu et al., 2024)). Tables 1 and 2 present the comparative results on MIMIC-III and MIMIC-IV datasets respectively. Statistical significance was evaluated using 10 rounds of bootstrapping, confirming these improvements are statistically significant ($p<0.05$) across all metrics.

Table 1: Performance Comparison on MIMIC-III

Methods	Jaccard	F1	DDI	Avg. # of Drugs
RETAIN	0.4887 ± 0.0028	0.6481 ± 0.0020	0.0835 ± 0.0020	20.4051 ± 0.2832
LEAP	0.4521 ± 0.0024	0.6138 ± 0.0026	0.0731 ± 0.0008	18.7138 ± 0.0666
GAMENet	0.5024 ± 0.0010	0.6595 ± 0.0008	0.0864 ± 0.0006	27.2145 ± 0.1141
SafeDrug	0.5054 ± 0.0024	0.6621 ± 0.0021	0.0589 ± 0.0005	19.9178 ± 0.1604
COGNet	0.5336 ± 0.0011	0.6869 ± 0.0010	0.0852 ± 0.0005	28.0900 ± 0.0950
REFINE	0.5235 ± 0.0018	0.6794 ± 0.0017	0.0393 ± 0.0002	–
ACDNet	0.5433 ± 0.0027	0.6957 ± 0.0021	0.0859 ± 0.0010	20.4900 ± 0.1197
LAMRec	0.4796 ± 0.0037	0.6360 ± 0.0030	0.0616 ± 0.0020	30.9211 ± 1.7195
LEADER	0.5179 ± 0.0021	0.6731 ± 0.0020	–	–
AMGNet	0.5416 ± 0.0023	0.6887 ± 0.0027	0.0590 ± 0.0011	20.2851 ± 0.1244
TRIXMED	**0.6922 ± 0.0087**	**0.8043 ± 0.0067**	**0.0420 ± 0.0005**	22.1900 ± 0.0022

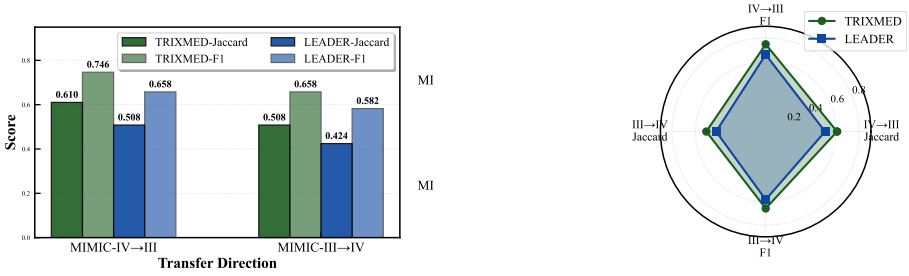
Table 2: Performance Comparison on MIMIC-IV

Methods	Jaccard	F1	DDI	Avg. # of Drugs
RETAIN	0.4152 ± 0.0006	0.5688 ± 0.0043	0.0939 ± 0.0015	10.8602 ± 0.0736
LEAP	0.4287 ± 0.0012	0.5820 ± 0.0012	0.0592 ± 0.0004	11.5198 ± 0.0459
GAMENet	0.4336 ± 0.0082	0.5871 ± 0.0030	0.0890 ± 0.0003	18.4426 ± 0.0474
SafeDrug	0.4295 ± 0.0027	0.5820 ± 0.0024	0.0740 ± 0.0004	14.4705 ± 0.0575
COGNet	0.4884 ± 0.0009	0.6367 ± 0.0009	0.0894 ± 0.0003	19.7235 ± 0.0242
REFINE	0.4620 ± 0.0003	0.6350 ± 0.0003	0.0430 ± 0.0001	–
ACDNet	0.5077 ± 0.0015	0.6564 ± 0.0013	0.0849 ± 0.0005	12.7024 ± 0.0005
LAMRec	0.4683 ± 0.0011	0.6198 ± 0.0013	0.0615 ± 0.0016	26.9675 ± 1.9922
LEADER	0.4779 ± 0.0021	0.6296 ± 0.0020	–	–
AMGNet	0.5063 ± 0.0015	0.6491 ± 0.0012	0.0643 ± 0.0006	16.5753 ± 0.0747
TRIXMED	**0.5212 ± 0.0077**	**0.6669 ± 0.0068**	**0.0439 ± 0.0004**	21.8600 ± 0.0024

The performance comparison in demonstrates substantial improvements of TRIXMED over State-of-the-Art (SOTA) baselines across both datasets. On MIMIC-III, TRIXMED achieves a Jaccard similarity of 27.4% relative improvement over the strongest baseline ACDNet. The F1 score surpasses all baselines by substantial margins, with the closest competitor ACDNet. Notably, TRIXMED maintains this superior performance while recommending an average of 22.19 medications per visit, demonstrating its ability to provide comprehensive treatment plans without compromising safety. On the more challenging MIMIC-IV dataset, TRIXMED continues to outperform all baselines, representing relative improvements of 2.66% and 1.6% over the best competitor ACDNet.

4.3 Cross-Dataset Generalization

To rigorously evaluate the generalization capability of TRIXMED, we conducted comprehensive zero-shot transfer experiments between MIMIC-III and MIMIC-IV datasets. This evaluation paradigm tests the model's ability to generalize across different medical datasets without any fine-tuning, providing insights into real-world deployment scenarios where models must adapt to new clinical environments. Our experimental setup involved bidirectional transfer learning: (1) models trained on MIMIC-IV and evaluated on MIMIC-III (IV→III), and (2) models trained on MIMIC-III and evaluated on MIMIC-IV (III→IV).



(a) Cross-Dataset Transfer Performance

(b) Improvement Analysis

Figure 2: Zero-shot Transfer Learning Performance Comparison

As illustrated in Figure 2(a), in the IV→III direction, TRIXMED represents a 20.07% improvement in Jaccard similarity and 13.37% improvement in F1 score. The III→IV transfer shows similar advantages. The superior IV→III performance suggests that more comprehensive of MIMIC-IV and recent data provides richer representations that generalize well to earlier clinical patterns. Conversely, the III→IV transfer faces the additional challenge of adapting to evolved clinical practices and potentially new medication protocols.

Figure 2(b) visualizes the relative improvements, highlighting that advantage of TRIXMED is most pronounced in the more challenging scenarios. The consistent performance gaps across both metrics and transfer directions validate that MoE architecture with clustering-based routing provides more robust and generalizable representations compared to monolithic approaches.

4.4 Ablation Study

Model Architecture To systematically evaluate the contribution of each component in TRIXMED, we conducted comprehensive ablation experiments on the MIMIC-III dataset. Table 3 presents the results of three model configurations: (1) w/o MoE: baseline model without the mixture-of-experts architecture, (2) Std. MoE: standard MoE without clustering-based routing, and (3) TRIXMED: our model with both MoE and clustering-guided routing.

The results clearly demonstrate that both components are critical to the overall performance of model. The MoE framework offers the architectural foundation for specialized learning, while the clustering-based routing mechanism guarantees the efficient exploitation of this capacity through intelligent patient–expert assignment. It is precisely this synergy that enables TRIXMED to achieve SOTA performance in medication recommendation.

Table 3: Ablation study on MIMIC-III dataset

Models	Jaccard	Precision	Recall	F1
w/o MoE	0.6897	0.7731	0.8789	0.8021
Std. MoE	0.6725	0.79278	0.8430	0.7835
TRIXMED	0.6922	0.7701	0.8858	0.804625

Hyperparameter Sensitivity Analysis To determine the optimal configuration for TRIXMED, we conducted extensive experiments varying critical hyperparameters: the number of clusters and the top-n routing value. Figure 3 presents comprehensive results across six evaluation metrics, revealing important insights about model behavior.

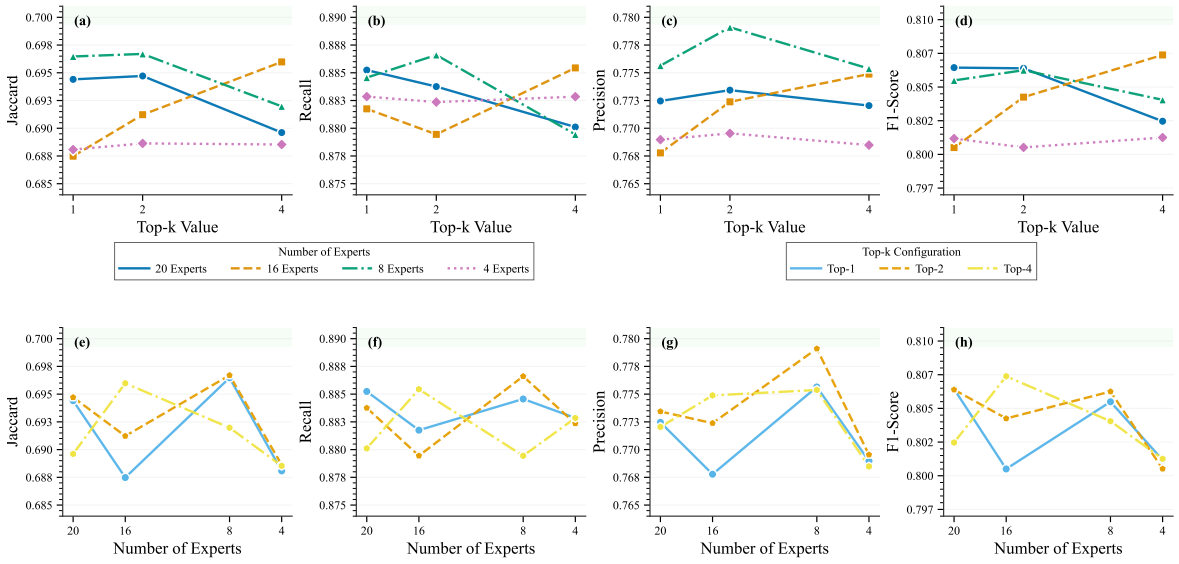


Figure 3: Impact of Cluster Number and top-n Routing on Model Performance

Figures 3(a-d) demonstrate how varying the number of experts (20, 16, 8, 4) affects model performance across different top-n values. Figures 3(e-h) illustrate how top-n (4, 2, 1) routing influences performance for different expert configurations. We observe optimal performance with 8 clusters, achieving a highest score at top-n=2, which gradually decreases as top-n increases. This suggests that moderate cluster granularity provides the best balance between patient stratification and maintaining sufficient training data per expert. Based on comprehensive analysis, we identify K=8 clusters with top-n=2 as the optimal configuration for TRIXMED.

4.5 Expert Sensitivity Analysis

To validate the clinical coherence of our triage-based routing mechanism, we present the analysis results using four typical patient clusters as examples (Figure 4). The analysis demonstrates that the effectiveness of our approach in automatically grouping patients with similar clinical profiles. The analysis reveals that our clustering algorithm

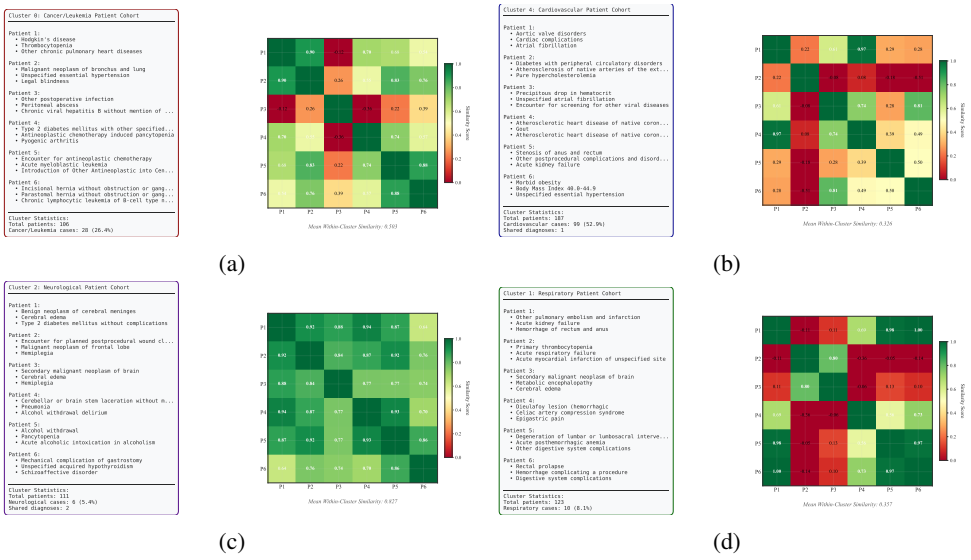


Figure 4: Patient grouping of similar medical conditions

successfully identifies four distinct patient cohorts: cancer/leukemia patients (Cluster 0, 26.4% cancer cases), cardiovascular patients (Cluster 4, 52.9% cardiovascular cases), neurological patients (Cluster 2, 5.4% neurological cases), and respiratory patients (Cluster 1, 8.1% respiratory cases). Each cluster exhibits high within-group similarity as evidenced by the similarity matrices, with neurological patients demonstrating the highest internal coherence (mean similarity 0.870), followed by cancer patients (0.440), suggesting these conditions manifest in more homogeneous clinical patterns. The cardiovascular and respiratory clusters show moderate similarities (0.333 and 0.269 respectively), reflecting the greater heterogeneity in these disease presentations.

The clinical validity of our approach is further substantiated by the semantic coherence within each cluster. For instance, Cluster 0 aggregates patients with various malignancies and associated complications such as chemotherapy-induced pancytopenia and antineoplastic treatment protocols, while Cluster 4 consolidates cardiovascular conditions including aortic valve disorders, atrial fibrillation, and atherosclerotic diseases. This automatic disease-specific grouping ensures that each expert network receives training data from a clinically homogeneous patient population, enabling specialized learning of treatment patterns. This patient stratification mechanism represents a significant advancement over traditional one-size-fits-all approaches, as it enables TRIXMED to develop targeted therapeutic expertise for distinct patient phenotypes while maintaining transparent routing decisions.

4.6 Efficiency Evaluation

TRIXMED demonstrates competitive efficiency, achieving an inference speed of 15.59 seconds per iteration on the MIMIC-III dataset. While this is slower than LEADER, it represents a substantial improvement over traditional deep learning approaches such as LEAP and GAMENet. The memory consumption of TRIXMED is 18.18 GB, marginally higher than that of LEADER, which is a reasonable trade-off given the incorporation of MoE components that enhance model capacity and clinical adaptability.

Table 4: Performance and inference speed of TRIXMED

Datasets	Inference Speed	Memory Consumption
MIMIC-III	15.59 s/iter	18.18 GB
LEADER	4.485 s/iter	17 GB
LEAP	32.31 s/iter	–
GAMENet	26.85 s/iter	–

5 Conclusion

In this paper, we propose TRIXMED for personalized drug recommendation. Our model extracts and encodes patient heterogeneity information through clustering-based stratification to augment the medication recommendation task. We design a triage-based routing mechanism to provide additional flexibility and interpretability. We evaluated TRIXMED using benchmark data and showed better accuracy and efficiency.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

Suman Bhoi, Mong Li Lee, Wynne Hsu, and Ngiap Chuan Tan. Refine: A fine-grained medication recommendation system using deep learning and personalized drug interaction modeling. *Advances in Neural Information Processing Systems*, 36:24013–24024, 2023.

Lei Chen, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou. Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS one*, 7 (4):e35254, 2012.

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

Basma M Hassan and Shahd Mohamed Elagamy. Personalized medical recommendation system with machine learning. *Neural Computing and Applications*, 37(9):6431–6447, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

J Jayapradha, Yukta Kulkarni, Palanichamy Naveen, Elham Abdulwahab Anaam, et al. Treatment recommendation using bert personalization. *Journal of Informatics and Web Engineering*, 3(3):41–62, 2024.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Shiji Li, Haitao Wang, Jianfeng He, and Xing Chen. Amgnet: An attention-guided multi-graph collaborative decision network for safe medication recommendation. *Electronics*, 14(4):760, 2025.

Xiang Li, Shunpan Liang, Yulei Hou, and Tengfei Ma. Stratmed: Relevance stratification between biomedical entities for sparsity on medication recommendation. *Knowledge-Based Systems*, 284:111239, 2024.

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

Xuan Lu, Yanhong Hao, Furong Peng, Zheqing Zhu, and Zhanwen Cheng. Expdrug: An explainable drug recommendation model based on space feature mapping. *Neurocomputing*, 619:129021, 2025.

Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

Jiacong Mi, Yi Zu, Zhuoyuan Wang, and Jieyue He. Acdnet: Attention-guided collaborative decision network for effective medication recommendation. *Journal of Biomedical Informatics*, 149:104570, 2024.

Luis Fernando Granda Morales, Priscila Valdiviezo-Diaz, Ruth Reátegui, and Luis Barba-Guaman. Drug recommendation system for diabetes using a collaborative filtering and clustering approach: Development and performance evaluation. *J Med Internet Res*, 24(7):e37233, 2022.

Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, 43(11):1130–1139, 2005.

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009(1):421425, 2009.

Yaqi Su, Yuliang Shi, Wu Lee, Lin Cheng, and Hongmei Guo. Tahdnet: Time-aware hierarchical dependency network for medication recommendation. *Journal of Biomedical Informatics*, 129:104069, 2022.

Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3970–3980, 2022.

Yunsen Tang, Ning Liu, Haitao Yuan, Yonghe Yan, Lei Liu, Weixing Tan, and Lizhen Cui. Lamrec: Label-aware multi-view drug recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 2230–2239, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679656.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

U.S. Food and Drug Administration. Office of pharmaceutical quality 2023 annual report, 2023. URL https://www.fda.gov. Accessed: [2025.3.18].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

World Health Organization. Influenza (seasonal), 2023. URL https://www.who.int/zh/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: 2024-03-15.

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

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*, 2021.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10665–10673, 2021.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 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*, pp. 1315–1324, 2017.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

A Appendix

A.1 Training Details

The implementation code is available in the GitHub repository[1]. Structured EHRs from `data/mimic3/output.csv` serves as the input, with model checkpoints and training artifacts stored in `/output/class_router_mimic-III`. To optimize computational efficiency, we employ gradient accumulation over 64 steps combined with a per-device batch size of 1, effectively simulating a larger batch size. Training spans 3 epochs with a learning rate of 1×10^{-4}, stabilized by 100 warmup steps. Logging intervals are set to every 10 steps, and model checkpoints are saved at the end of each epoch. Memory optimization is achieved through bf16 mixed-precision training and gradient checkpointing. Training metrics are tracked via TensorBoard under `/denseMOE/logs`, with gradient norms clipped at 0.3 to mitigate instability during optimization.

[1] https://anonymous.4open.science/r/TRIXMED-0092

Hyperparameter settings The configuration integrates the LlaMA2-7B architecture with a causal language modeling objective. Optimization relies on the AdamW(Yao et al., 2021) algorithm with a fixed learning rate of 1×10^{-4} and linear warmup. Regularization is enforced through gradient clipping at a maximum norm of 0.3. Hardware constraints are addressed via bf16 mixed-precision training and gradient checkpointing, balancing computational demands with model performance. The effective batch size of 64 is maintained through gradient accumulation. Training progress is monitored through frequent metric logging (every 10 steps), and model states are preserved at epoch boundaries. This setup adheres to established practices for fine-tuning LLMs on clinical datasets, prioritizing stability and reproducibility.

Algorithm 1 TRIXMED: Training Procedure for Triage-Based Drug Recommendation

Require: Training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, warm-up ratio $\rho = 0.3$, number of clusters K, number of experts E, top-k routing parameter
Ensure: Trained TRIXMED model with experts $\mathcal{E} = \{E_1, \ldots, E_K\}$, routing weights $\mathbf{W}_e$

Phase 1: Warm-up Feature Learning
Sample warm-up subset $\mathcal{D}_{\text{warm}} \subset \mathcal{D}$ with $|\mathcal{D}_{\text{warm}}| = \rho \cdot N$
Initialize LLaMA3.1-8B with LoRA adapters as f_{warm}
for epoch $= 1$ to T_{warm} **do**
  for batch $(\mathbf{x}_b, y_b) \in \mathcal{D}_{\text{warm}}$ **do**
    Compute embeddings: $e_b = f_{\text{embed}}(\mathbf{x}_b)$ using $\mathbf{E}_d, \mathbf{E}_p, \mathbf{E}_m$
    Forward pass: $\hat{y}_b = f_{\text{warm}}(e_b; \theta)$
    Compute loss: $\mathcal{L}_{\text{warm}} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{med}}\mathcal{L}_{\text{BCE}}(y_b, \hat{y}_b)$
    Update parameters: $\theta \leftarrow \theta - \eta_{\text{warm}}\nabla_\theta \mathcal{L}_{\text{warm}}$
  end for
end for
Phase 2: Feature Extraction and Patient Stratification
Freeze warm-up model parameters θ^*
for patient $i = 1$ to N **do**
  Extract features: $\mathbf{H}_i = f_{\text{warm}}(f_{\text{embed}}(\mathbf{x}_i); \theta^*)$
  Apply mean pooling: $\bar{h}_i = \frac{1}{T}\sum_{t=1}^{T} h_i^{(t)}$
  Apply PCA: $z_i = \mathbf{W}_{\text{pca}} \cdot \bar{h}_i$
  Standardize: $\tilde{z}_i = (z_i - \mu_z)/(\sigma_z + \epsilon)$
end for
Apply K-means: $\mathcal{G} = \text{K-means}(\{\tilde{z}_i\}_{i=1}^{N}, K)$
Obtain cluster assignments: $\{c_i\}_{i=1}^{N}$ where $c_i \in \{1, \ldots, K\}$
Phase 3: Expert Training with Triage Routing
Initialize experts $\{E_1, \ldots, E_K\}$ from f_{warm}
Initialize routing weights $\mathbf{W}_e \in \mathbb{R}^{K \times E}$, $\mathbf{b}_e \in \mathbb{R}^E$
Partition data: $\mathcal{D}_k = \{(\mathbf{x}_i, y_i) | c_i = k\}$ for $k = 1, \ldots, K$
for epoch $= 1$ to T_{expert} **do**
  for batch $(\mathbf{x}_b, y_b, \mathbf{c}_b)$ from $\mathcal{D}_{\text{train}}$ **do**
    Compute expert affinities: $\mathbf{Z} = \text{OneHot}(\mathbf{c}_b)\mathbf{W}_e + \mathbf{b}_e$
    Normalize: $\mathbf{G} = \text{Softmax}(\mathbf{Z})$
    Select top-k: $(\mathbf{\Gamma}, \mathbf{I}) = \text{Top\_n}(\mathbf{G}, k)$
    for each expert j in activated experts **do**
      Compute mask: $\mathbf{M}_j = \mathbb{I}(\mathbf{I} = j)$
      Expert forward: $\mathbf{H}_{\text{expert}}^{(j)} = E_j(\mathbf{H}_{\text{in}} \odot \mathbf{M}_j)$
    end for
    Synthesize outputs: $\mathbf{H}_{\text{final}} = \sum_{j=1}^{k} \mathbf{\Gamma}[:, j] \cdot \mathbf{H}_{\text{expert}}^{(\mathbf{I}[:,j])}$
    Transform to predictions: $\hat{m}_b^{(T)} = \sigma(\mathbf{W}_o\mathbf{H}_{\text{final}} + \mathbf{b}_o)$
    Compute auxiliary loss: $\mathcal{L}_{\text{aux}} = E\sum_{e=1}^{E} f_e \cdot r_e$
    Compute total loss: $\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{BCE}}(m_b^{(T)}, \hat{m}_b^{(T)}) + \lambda\mathcal{L}_{\text{aux}}$
    Update experts and routing: $\{\theta_{E_k}, \mathbf{W}_e\} \leftarrow \{\theta_{E_k}, \mathbf{W}_e\} - \eta_{\text{expert}}\nabla\mathcal{L}_{\text{final}}$
  end for
end for
return Trained model $\{\mathcal{E}, \mathbf{W}_e, \mathbf{W}_o, \mathbf{b}_o\}$

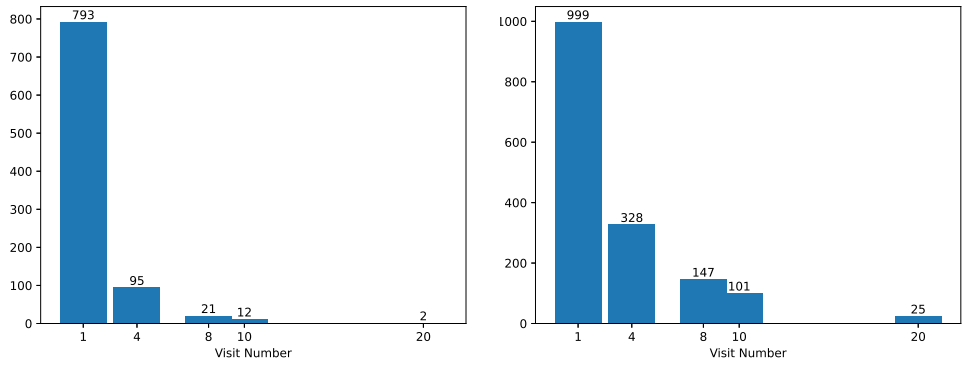
Algorithm We summarize the procedure of training TRIXMED in Algorithm 1. The algorithm comprises three distinct phases. In Phase 1 (Lines 2-8), we perform warm-up feature learning on a subset of the training data to learn initial patient representations. We sample 30% of the training set and train a model with LoRA adapters to capture fundamental clinical patterns. In Phase 2 (Lines 10-15), we extract comprehensive features from the entire

dataset using the frozen warm-up model and apply K-means clustering to identify K distinct patient subgroups based on their clinical characteristics. Each patient is assigned a cluster label that represents their phenotypic group. In Phase 3 (Lines 17-28), we initialize K expert networks from the warm-up model and train them using our triage-based routing mechanism. For each training batch, we compute cluster-based expert affinities (Line 22) and select the top-n most relevant experts (Line 24). The selected experts process their assigned patients (Lines 25-27), and their outputs are synthesized based on routing weights (Line 28). We employ an auxiliary load balancing loss (Line 30) to prevent routing collapse and ensure balanced expert utilization. The total loss combines the medication prediction loss with this balancing term (Line 31), enabling joint optimization of both prediction accuracy and expert specialization.

A.2 Integrated Description of Data Processing Workflow

The data processing pipeline comprises four sequential stages. First, data loading and preprocessing are performed, where multiple CSV files including diagnostic records, surgical procedures, prescriptions, and admission data are securely loaded via exception-handling mechanisms.

As shown in Figure 5, the dataset exhibits distinct visit frequency distributions. In MIMIC-III (Figure 5a), 793 patients had only 1 visit, with a sharp decline in counts for higher frequencies (e.g., 21 patients with 8 visits, 2 patients with 20 visits). MIMIC-IV (Johnson et al., 2020) (Figure 5b) demonstrates a similar long-tail pattern but with larger populations, including 999 single-visit patients and 25 patients with 20 visits.



(a) (b)

Figure 5: Patient Visit Distribution in MIMIC-III and MIMIC-IV

During this phase, mapping dictionaries are constructed to convert ICD9/10 codes to diagnosis/surgery names and ATC codes to medication names. Medication names are standardized by matching them to predefined ATC codes or randomly selecting common ATC codes if no direct match is found. Subsequently, patient data grouping organizes records by patient and admission IDs, categorizing diagnoses, procedures, and medications into structured formats while integrating demographic metadata (e.g., insurance status, ethnicity). Next, the SFT dataset generation phase filters patients with admissions, transforms historical clinical information (diagnoses, procedures, medications) into natural language descriptions, and combines these with current visit details to form structured samples. Each sample includes an instruction, input (historical and current visit context), output (recommended medications with ATC codes and names), and metadata, tailored for medication recommendation tasks. Finally, the processed data are saved into three outputs: the full dataset (mimic3_full_dataset.csv), a simplified version retaining only instruction-input-output fields (mimic3_train_dataset.csv), and a validation subset (mimic3_test_dataset.csv) to ensure compatibility with supervised fine-tuning frameworks.

This workflow ensures robust data standardization, temporal context preservation, and task-specific formatting for downstream model training.

```
Instruction: Based on the patient's historical medical records, predict the medications needed for the
↪  current visit.
Input:
  Patient History:
    [Details of historical visit 1]
    ...
  Current Visit:
    [Details of current visit]
Output:
  Recommended medications:
    [ATC code 1 (medication name 1),ATC code 2 (medication name 2),...]
```

A.3 SINGLE-VISIT PERFORMANCE

Table 5 presents a comprehensive comparison of single-visit performance across multiple methods on the MIMIC-III and MIMIC-IV datasets, evaluated using Jaccard and F1 scores. Notably, TRIXMED achieves SOTA results on MIMIC-III, highlighting superior capability in of TRIXMED balancing medication set overlap and predictive robustness. On MIMIC-IV, TRIXMED maintains competitive performance, demonstrating superior capability in addressing structural dataset discrepancies when generalizing to more complex clinical patterns within MIMIC-IV.

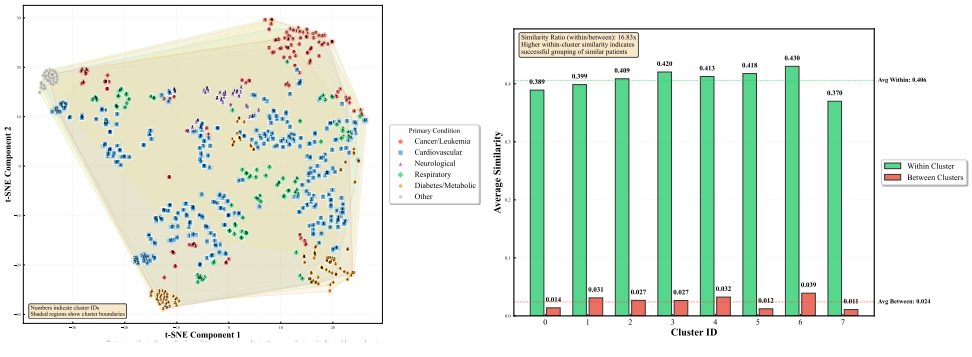
Table 5: Single-visit Performance Comparison on MIMIC Dataset

Methods	MIMIC-III		MIMIC-IV	
	Jaccard	F1	Jaccard	F1
RETAIN	0.4811 ± 0.0053	0.6403 ± 0.0049	0.4165 ± 0.0035	0.5707 ± 0.0042
GAMENet	0.4840 ± 0.0038	0.6442 ± 0.0036	0.4292 ± 0.0041	0.5819 ± 0.0040
SafeDrug	0.4900 ± 0.0043	0.6481 ± 0.0042	0.4214 ± 0.0073	0.5749 ± 0.0073
COGNet	0.5336 ± 0.0011	0.6869 ± 0.0010	0.4884 ± 0.0009	0.6367 ± 0.0009
LEADER	0.5179 ± 0.0021	0.6731 ± 0.0020	0.4539 ± 0.0026	0.6071 ± 0.0026
TRIXMED	$\mathbf{0.6823 \pm 0.0024}$	$\mathbf{0.7715 \pm 0.0014}$	$\mathbf{0.5637 \pm 0.0084}$	$\mathbf{0.7071 \pm 0.0017}$

In summary, the TRIXMED architecture demonstrates critical value in addressing longitudinal clinical data analysis tasks, with its core advantage lying in its systematic approach to resolving multidimensional clinical challenges. The model not only achieves an effective balance between precision and coverage but also significantly enhances generalization performance for heterogeneous medical data through its inherent adaptive mechanisms. The performance variations observed across different datasets further validate that intelligent decision-making models for real-world clinical scenarios require dynamic adaptation capabilities to cope with continuously evolving medical documentation standards and drug interaction patterns. By introducing structural innovations, TRIXMED provides a forward-looking solution to these challenges, offering an important technical pathway for the evolution of clinical decision support systems.

A.4 EXPERT ROUTING ANALYSIS

Figure 6 clearly demonstrates distinct patient clustering and expert assignment patterns through three complementary visualizations, offering robust support for its capacity to automatically identify clinically similar patients and assign them to domain-specific experts, thereby effectively simulating the clinical triage process.



(a) (b)

Figure 6: patient similarity clusterin

The t-SNE visualization in Figure 6 (left) reveals that clustering mechanism successfully groups patients with similar medical conditions into distinct clusters. Each cluster, denoted by numerical IDs (0-7), demonstrates clear spatial separation with well-defined boundaries (shown as shaded regions). Notably, patients with cancer/leukemia (red markers) predominantly concentrate in clusters 0 and 3, while cardiovascular patients (blue squares) form dense groups in clusters 1, 4, and 5. This natural segregation validates that our clustering approach captures genuine clinical phenotypes rather than arbitrary divisions.

The right panel of Figure 6 provides quantitative evidence of clustering effectiveness. Within-cluster similarity scores have mean 0.41, significantly exceeding between-cluster similarities of mean 0.025. This 16.4× ratio demonstrates that patients within the same cluster share substantially more clinical features than those in different clusters. Cluster 6 exhibits the highest internal coherence (0.43), suggesting it captures a particularly homogeneous patient subgroup, while Cluster 7 shows slightly lower but still strong cohesion (0.37), potentially representing a more diverse but clinically related population.

The clustering-based routing ensures that when a new patient enters the system, they are automatically matched to the expert cluster with the most similar historical cases. Healthcare providers can examine which historical patients influenced an expert's training and understand why specific experts were selected for a given patient. The clear condition-cluster associations (e.g., Cluster 0 for cardiovascular-cancer combinations, Cluster 7 for diverse presentations) provide transparency that is essential for clinical adoption.

A.5 LLM USE STATEMENT

We hereby declare that the use of LLMs in this study was strictly confined to enhancing the clarity and readability of the text. All core ideas, research design, experimental methodology, data analysis, and scientific conclusions presented in this paper were independently developed by the authors without any substantive contribution from LLMs. We extend our sincere gratitude to the reviewers for their time and valuable feedback on our work.