TreatRAG: A Framework for Personalized Treatment Recommendation

Chao-Chin Liu
Georgetown University
Washington DC, USA
cl1477@georgetown.edu

Hao-Ren Yao
Carnegie Mellon University
Pittsburgh, USA
haoreny@cs.cmu.edu

Der-Chen Chang
Georgetown University
Washington DC, USA
chang@georgetown.edu

Ophir Frieder
Georgetown University
Washington DC, USA
ophir@ir.cs.georgetown.edu

Abstract

Medication recommendation is a critical function of clinical decision support systems, directly influencing patient safety and treatment efficacy. While large language models (LLMs) show promise in clinical tasks such as summarization and question answering, their ability to make accurate treatment predictions remains limited, in part, due to their lack of specialized medical knowledge and exposure to real-world patient data. We introduce **TreatRAG**, an interpretable, model-agnostic retrieval-augmented generation (RAG) framework aimed at early-stage development to enhance medication recommendation accuracy using publicly available clinical data; thus, TreatRAG forms a critical foundational step toward future clinical validation and domain expert involvement. TreatRAG retrieves similar patient cases, i.e., so called "digital twins", using interpretable N-gram Jaccard similarity and augments the input prompt to ground LLM predictions in real clinical scenarios. We evaluate our framework on the MIMIC-IV dataset using BioGPT, BioMistral, Phi3, and Flan-T5. TreatRAG-enhanced BioGPT improves its F1-score from 0.14 to 0.34, BioMistral from 0.22 to 0.54, Phi-3 from 0.09 to 0.16, and Flan-T5 from 0.23 to 0.30, while also lowering, often significantly, the hallucination rate. Our model-agnostic framework offers a flexible, effective, and interpretable solution to advance the reliability of LLMs in clinical decision support.

CCS Concepts

• **Applied computing** → **Health informatics**.

Keywords

Retrieval-Augmented Generation, Medical Digital Twins, Medication Recommendation, Large Language Models

ACM Reference Format:
Chao-Chin Liu, Hao-Ren Yao, Der-Chen Chang, and Ophir Frieder. 2025. TreatRAG: A Framework for Personalized Treatment Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3705328.3748022

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748022

1 Introduction

Clinical decision support (CDS) systems assist clinicians in making accurate and timely decisions [3, 15]. Among these applications, medication recommendation is particularly critical, as medication recommendation directly impacts patient safety and treatment effectiveness [10, 18]. For instance, a CDS system can analyze a patient's medical history, allergies, and current medications to suggest the most appropriate antibiotic while flagging potential drug interactions. Thus, medication recommendation systems reduce medical errors and improve patient outcomes [21]. Large Language Models (LLMs) have demonstrated promising results on diverse healthcare tasks. In the medical field, LLMs were applied to tasks such as clinical note summarization[6, 23] and question-answering [2, 4]. While LLMs excel at extracting and organizing key information from patient records, they still face challenges in making highly accurate clinical decisions.

We introduce TreatRAG, a Retrieval-Augmented Generation (RAG) framework for patient medication recommendations, designed to enhance LLM reasoning by integrating real-world patient data, e.g., MIMIC-IV [9], without requiring additional training. Incorporating clinical data ensures that recommendations are informed by up-to-date, relevant clinical information. Unlike approaches that rely solely on pre-trained LLM knowledge, our method incorporates de-identified Electronic Health Records (EHRs) directly into the process, improving the precision of medication suggestions. Our framework bridges the gap between generic LLM outputs and clinically actionable insights by combining retrieval-augmented generation with structured EHR analysis.

To associate patient records with relevant historical cases, we selected N-gram Jaccard similarity due to its interpretability and simplicity, qualities highly beneficial in clinical settings, and validated its effectiveness through an ablation study comparing it to dense embedding methods (e.g., SBERT with FAISS). This method identifies clinically relevant cases by comparing a patient's history and diagnoses with past records. Gap statistics filter out less informative matches, retaining only contextually valuable cases. Retrieved cases are then structured into prompts for the LLM, which generate medication recommendations based on both patient data and historical

treatments. We implement our model-agnostic RAG framework using BioGPT[13], BioMistral[11], Flan-T5 [7] and Phi3[1]. Our main contributions are as follows:

- We propose a RAG framework that integrates real patient data as input to enhance the clinical actionability of pre-trained LLMs without additional training costs.
- We demonstrate the flexibility of our RAG framework by implementing and evaluating it with BioGPT, BioMistral, Flan-T5, and Phi-3, incorporating the MIMIC-IV dataset in our experiments.

2 Related Work

Advances in deep learning greatly improved EHR representation learning, enabling a wide range of CDS tasks [5, 20, 27–30]. To enhance interpretability, safety, and efficacy , models employ sequential modeling [5], reinforcement learning [31], and graph-based approaches [20, 27], where the latter integrating drug-drug interaction (DDI) graphs, to maximize treatment safety and effectiveness. Language modeling via transformers, using the pre-train, fine-tune paradigm, greatly improves the accuracy and fairness of medication recommendations [19, 26], especially on rare diseases [32].

While recent models have advanced medication recommendations, challenges remain in adapting to the diverse patient trajectories. LLMs offer improved generalization and robustness by leveraging broad medical knowledge and flexible reasoning capabilities [12, 24]. However, hallucinations and contextually irrelevant output often undermine their utility in clinical settings [17]. RAG addresses the aforemention limitations by incorporating external, domain-specific information, such as medical knowledge or patient histories, into the inference process, thereby enhancing the contextual accuracy and reliability of LLM-generated responses [8, 17, 25]. Our approach integrates similar patient trajectories as retrieval context to ground LLM-based medication recommendations in real-world clinical patterns, resulting in more accurate and robust decisions.

3 Methodology

TreatRAG operates in three stages: (1) transforming structured clinical records into interpretable text representations, (2) retrieving similar patient cases, commonly referred to as "digital twins" based on textual similarity, and (3) constructing prompts that integrate the retrieved cases with the target patient's information to guide generation, as illustrated in Figure 1. TreatRAG adopts a transparent retrieval mechanism based on N-gram Jaccard similarity. Given two patient case texts, represented as sets of N-grams, the Jaccard similarity is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B are the N-gram sets for two patient cases. This token-based approach enables interpretable case matching that aligns with the structure and semantics of clinical narratives. By grounding recommendations in retrieved examples, the framework enhances the factual reliability of language models without requiring model retraining or fine-tuning.

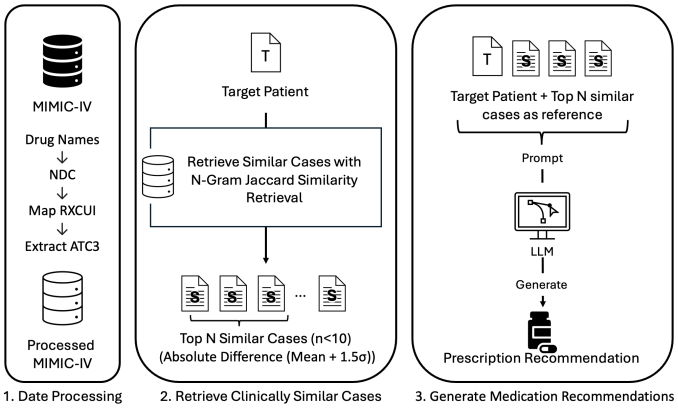

Figure 1: Overview of the TreatRAG framework. The system consists of three main stages: (1) Data Processing: transforming structured EHR data into natural language prompts, (2) Retrieve Clinically Similar Cases: retrieving similar patient cases using N-gram Jaccard similarity, and (3) Generate Medication Recommendation: generating medication recommendations using a pretrained LLM. Retrieved cases are incorporated into the prompt to ground predictions in real-world clinical examples. NDC: National Drug Code, RXUCI: RxNorm Concept Unique Identifier, ATC3: Level 3 Anatomical Therapeutic Chemical.

3.1 Medical Recommendation

The prompt structure is inspired by prior LLM-based healthcare work [12], which formats structured clinical inputs into natural language to enhance model compatibility. We preprocess each patient record into a structured textual prompt to ensure compatibility with LLMs. For each patient, we extract all visits in historical order and summarize each visit with its respective diagnoses and prescribed medications. The summaries are formatted into natural language using a consistent prompt template, as shown in Figure 2. Earlier visits include both diagnoses and prescriptions, while the final visit contains only diagnoses. The prompt ends with a cue: "Then, the patient should be prescribed:" to instruct the model to generate the appropriate medications for the most recent visit.

The patient has < TOTAL_VISIT_NUM > times ICU visits. In the first visit, the patient had diagnosis: < DIAGNOSIS >, ..., < DIAGNOSIS >. The patient was prescribed: <PRESCRIPTION>, ..., < PRESCRIPTION >. In the second visit, ... In this visit, the patient has diagnosis: < DIAGNOSIS>, ..., < DIAGNOSIS >. Then, the patient should be prescribed:

Figure 2: Prompt template for medication recommendation.

Medication prediction is explicitly conditioned on the diagnoses from the patient's latest visit. These diagnoses represent the "query" for our retrieval-augmented generation framework and encompass a broad range of clinical conditions common in the MIMIC-IV dataset, such as infections, cardiovascular diseases, respiratory issues, and metabolic disorders. We do not limit our evaluation to a fixed set of ailments; instead, we allow the model to generalize across the naturally occurring diagnostic distribution within

MIMIC-IV. This setting follows prior work on medication recommendation [12, 14, 22], where the prediction task involves inferring appropriate medications based on structured diagnosis inputs.

4 Experiment

We investigate the research question: *Can an integrated, model-agnostic RAG framework statistically significantly improve the accuracy of medication prediction of LLM architectures?*

4.1 Data Preprocessing

We utilize MIMIC-IV, a comprehensive clinical dataset containing patient records from the Intensive Care Unit (ICU), including patient admission, diagnoses, and prescription data. The preprocessing stage focuses on extracting and structuring relevant information from these records to ensure consistency in representation. This includes standardizing drug identifiers by mapping them to common terminologies and integrating external medical knowledge to account for potential drug-drug interactions. By normalizing medication data and structuring them in a unified format, Anatomical Therapeutic Chemical (ATC) Classification, we enable efficient retrieval and facilitate accurate predictions.

4.2 Retrieval Mechanism

To improve prediction reliability, we retrieve similar patient cases to provide additional clinical context. We use N-gram Jaccard similarity, which directly compares textual structure, capturing both semantic content and word order—key for clinical interpretability. Patient records are tokenized into N-grams, forming the basis for similarity computation. We rigorously evaluated our framework using patient-level splits, ensuring no overlap between test patients and retrieval pools, and conducted adaptive retrieval selecting up to k similar cases, typically fewer than 6 after filtering by adaptive thresholds (mean + 1.5×standard deviation). This threshold was experimentally determined to balance retrieval precision and coverage. To meet the model's token limitations, truncation strategies such as summarizing or omitting older visits were applied. If no cases pass this adaptive threshold, we skip reference to avoid introducing misleading context. This reduces noise, filters out weak matches, and ensures that only informative cases are included. By combining interpretable similarity with adaptive filtering and truncation strategies, our method enhances retrieval precision and minimizes hallucinations—boosting reliability without model fine-tuning.

4.3 LLM-Based Prediction

We evaluated the performance of four state-of-the-art language models, Flan-T5-small, BioGPT, BioMistral-7B, and Phi-3-mini on the task of medication prediction using our proposed framework. BioGPT is a biomedical-specific generative model pre-trained on PubMed articles, enabling fluent, domain-aware text generation. BioMistral specializes in biomedical and clinical content, leveraging scientific literature and structured EHR data to enhance medical understanding. Flan-T5 is an instruction-tuned model built for general-purpose reasoning and well-suited for structured clinical prompts. Phi-3, a compact instruction-following model optimized for efficiency, brings strong few-shot performance to a variety of healthcare-related NLP tasks. In the TreatRAG framework, the retrieved similar case data are incorporated into the target patient's prompt, ensuring that medication recommendations are informed by relevant past cases. This retrieval-enhanced approach allows LLMs to generate more reliable, evidence-based treatment suggestions, bridging the gap between pre-trained knowledge and real-world clinical applications.

4.4 Model Implementation

We implement our framework using state-of-the-art, open-weight LLM models. Each model is prompted with the retrieved patient data as a reference for the main question, allowing it to adapt to clinical decision-making tasks. The system is optimized to fully utilize available GPUs, ensuring fast and accurate similarity computation across large datasets. This methodology ensures that our system improves both accuracy and interpretability by grounding medication recommendations in real-world patient data while maintaining the flexibility to adapt to evolving medical knowledge.

4.5 Evaluation

We evaluate the performance of our framework using two key metrics: F1-score and Jaccard similarity. The F1-score provides a balanced measure of precision and recall for the predicted ATC codes, ensuring that both over-prediction and under-prediction are penalized appropriately. Jaccard similarity quantifies the overlap between predicted and ground-truth medication sets, offering an interpretable metric to assess set-level agreement.

5 Results

We augmented all baseline models using retrieval-based augmentation, leading to consistent performance improvements across all evaluated architectures. Table 1 summarizes the statistics of the MIMIC dataset. As shown in Table 2, TreatRAG-enhanced models achieved substantial gains over their baselines, with TreatRAG BioMistral achieving the highest overall performance with F1-score of 0.54 and Jaccard similarity of 0.37. All improvements were statistically significant according to paired t-tests ($p < 0.05$), highlighting the effectiveness of our augmentation approach.

Table 1: MIMIC Dataset Statistics. Summary of the MIMIC-IV dataset used in our experiments, including patient counts, admissions, and disease diversity.

Metric	Value
Number of patients	180,733
Total number of admissions	431,231
Average admissions per patient	2.39
Number of diseases	25,809

6 Discussion

Our results suggest that TreatRAG is beneficial across different model architectures. While performance varies by model, the framework consistently enhances medication prediction. This shows that

Table 2: Evaluation Results. * indicates the statistically significant improvements (i.e., Paired t-test with p<0.05)

Model	F1-score	Jaccard Similarity
Baseline BioGPT	0.14	0.07
Baseline BioMistral	0.23	0.13
Baseline Phi3	0.09	0.05
Baseline Flan-T5	0.23	0.12
TreatRAG BioGPT	0.34*	0.21*
TreatRAG BioMistral	0.54*	0.37*
TreatRAG Phi3	0.16*	0.09*
TreatRAG Flan-T5	0.30*	0.17*

the model-agnostic TreatRAG framework can be developed to integrate with multiple LLMs, provided that effective retrieval strategies and relevant clinical datasets are utilized. In addition to its accuracy, our approach is computationally efficient. Unlike training or fine-tuning large models, our retrieval-mechanism, prompt-engineered approach operates directly on existing patient records without requiring model updates or retraining. This makes the method lightweight and easily deployable in real-world settings, especially when computational resources are limited.

It is worth noting that the evaluated models were not trained explicitly with ATC or medication prediction tasks. Consequently, they may occasionally generate hallucinated outputs, including non-existent three-character codes." These errors highlight the limitations of using pre-trained LLMs alone for structured clinical tasks and emphasize the value of retrieval grounding to enhance reliability and clinical validity. We conducted a hallucination analysis by extracting predicted drug terms and comparing them against valid ATC-3 codes from the WHO classification [16]. Any unmatched term was considered a hallucination. We report both the overall hallucination rate and the per-patient hallucination percentage.

Table 3: Hallucination analysis under baseline and TreatRAG. TreatRAG significantly reduces over hallucinations for BioGPT (84.41 → 39.85%) and BioMistral (72.83 → 32.76%), with smaller gains for Phi-3 and Flan-T5. Overall, TreatRAG lowers hallucination rates and improves factual reliability.

Method	BioGPT	BioMistral	Phi-3	Flan-T5
Hallucination Rate (%)				
Baseline	84.41%	72.83%	86.18%	42.26%
TreatRAG	39.85%	32.76%	86.16%	37.66%
Average Hallucination per Patient (%)				
Baseline	70.37%	57.17%	69.63%	60.38%
TreatRAG	28.21%	17.50%	25.43%	58.65%
Patients with 0% / 100% Hallucination (%)				
Baseline	19.00 / 68.00%	41.07 / 57.14%	22.13 / 55.32%	39.36 / 60.34%
TreatRAG	43.06 / 20.74%	76.69 / 16.40%	71.92 / 21.32%	41.25 / 58.63%

As shown in Table 3, TreatRAG substantially reduced hallucinations for BioGPT and BioMistral (from 84.41% to 39.85% and 72.83% to 32.76%, respectively), aligning with their F1 and Jaccard score improvements. Phi-3 showed minimal overall change but improved at the patient level, while Flan-T5 exhibited modest gains. These results confirm that retrieval grounding enhances both predictive accuracy and factual reliability in medication recommendation.

Table 4: Prevalence of Rare Diagnoses and Affected Patients in the MIMIC-IV. This table summarizes the frequency of rare diagnostic codes and the number of affected patients in the dataset. A diagnosis is considered rare if it appears in the records of fewer than a threshold n patients. The left column defines these rarity thresholds (n), the middle column lists the number of distinct diagnoses that fall under each threshold, and the right column reports how many unique patients have such a rare condition. These statistics highlight the long-tail distribution of medical diagnoses and motivate using retrieval-augmented models like TreatRAG, which can handle sparse and infrequent data more effectively.

Threshold (n)	Diagnoses $< n$	Patients Affected
< 50	9,699	25,524
< 100	11,660	39,154
< 200	13,491	58,473

Table 5: Performance Comparison Between TreatRAG and Baseline Across Rare Diseases. This table reports F1 and Jaccard scores for different model backbones (BioGPT, BioMistral, Phi3, FlanT5) under various rarity thresholds (< 50, < 100, < 200). TreatRAG consistently outperforms the baseline across all settings, particularly under rarer diagnostic conditions, demonstrating its effectiveness in handling sparse and imbalanced medical data.

Baseline (F1 / Jaccard)				
Threshold (n)	**BioGPT**	**BioMistral**	**Phi3**	**FlanT5**
< 50	0.10 / 0.04	0.14 / 0.22	0.08 / 0.04	0.23 / 0.13
< 100	0.14 / 0.06	0.27 / 0.16	0.08 / 0.05	0.25 / 0.13
< 200	0.15 / 0.07	0.21 / 0.12	0.09 / 0.05	0.26 / 0.15
TreatRAG (F1 / Jaccard)				
Threshold (n)	**BioGPT**	**BioMistral**	**Phi3**	**FlanT5**
< 50	0.33 / 0.22	0.23 / 0.35	0.12 / 0.05	0.23 / 0.13
< 100	0.33 / 0.22	0.27 / 0.18	0.14 / 0.08	0.26 / 0.14
< 200	0.34 / 0.23	0.26 / 0.17	0.12 / 0.05	0.28 / 0.16

Rare diseases pose a significant challenge for clinical decision support systems due to the lack of sufficient data to train disease-specific models. In our evaluation, we define a diagnosis as rare if it appears in the records of fewer than n patients. To assess TreatRAG's performance under these low-resource conditions, Table 4 summarizes the statistic of rare diagnoses in the MIMIC-IV dataset. The table reports the number of distinct diagnostic codes that meet each rarity threshold, along with the total number of patients affected. These statistics reflect the long-tail nature of clinical

data and underscore the need for models that can effectively handle sparse and infrequent conditions.

As shown in Table 5, TreatRAG consistently outperforms baseline models across all rarity thresholds, with particularly strong improvements under the rarest settings (<50 patients). For example, BioGPT's F1-score more than triples, from 0.10 to 0.33, and BioMistral's Jaccard similarity improves from 0.22 to 0.35. These results highlight TreatRAG's ability to improve clinical reliability in data-sparse environments—an essential capability in rare disease scenarios, where annotated training data are often scarce. TreatRAG's zero-shot retrieval-based augmentation proves both practical and impactful in these contexts.

7 Ablation Studies

We conducted a comprehensive set of ablation experiments to evaluate the effect of different representation strategies and similarity functions on retrieval performance. We compared the following five retrieval configurations:

- **Full Sentence + Jaccard Similarity**: Each patient's full history is treated as raw text. Queries are compared using Jaccard similarity over word sets, capturing simple lexical overlap.
- **Full Sentence + Cosine Similarity**: Patient texts are transformed into TF-IDF vectors. Similarity is computed using cosine similarity, measuring the alignment of term-weighted representations.
- **DP Pair Embedding (Patient-level) + Vector Search**: Each visit is embedded as a diagnosis-prescription (DP) pair using a Sentence-BERT model (`all-mpnet-base-v2`). Visit embeddings are aggregated via mean pooling into a single patient-level vector. Retrieval is performed using cosine similarity (via FAISS with L2 distance on normalized vectors).
- **DP Pair Embedding (Visit-level) + Vector Search**: Visit embeddings are stored individually in the FAISS index, allowing fine-grained, visit-level matching. This configuration bypasses patient-level aggregation.
- **Full Sentence Embedding + Vector Search**: Patient's full history is encoded into a single dense vector using SBERT. Retrieval is performed using cosine similarity through FAISS.

Performance was evaluated using F1-score and Jaccard Score. Tables 6 show that Full Sent + Jaccard similarity consistently outperforms TF-IDF-based cosine similarity and dense vector retrieval across all models. Despite its simplicity, the token-level overlap captured by Jaccard similarity is highly effective for identifying relevant patient cases, yielding the highest F1-scores and Jaccard similarities. Among vector methods, visit-level DP pair embeddings perform competitively, particularly for BioMistral and Flan-T5. Full sentence embeddings provide reasonable baseline performance but are generally outperformed by structured DP-pair representations that better preserve clinical intent.

8 Future Work

A key direction for improvement is optimizing the retrieval mechanism. While our current method relies on structured retrieval from MIMIC-IV, future work can explore alternative strategies to enhance flexibility and performance. Incorporating more diverse medical datasets could further improve predictive power. Planned extensions include:

- External EHR datasets beyond MIMIC-IV to ensure broader generalization across different patient populations.
- Clinical guidelines and expert-annotated resources to ground recommendations in established medical knowledge.
- Pharmacogenomic data to personalize medication recommendations based on genetic factors.
- Incorporate clinical knowledge bases (e.g., UMLS, DrugBank) and expert reviews to ensure closer alignment with clinically safe and deployable systems.
- Integrating drug-drug interaction (DDI) detection into the retrieval and generation pipeline.

To enhance patient safety, we aim to extend our framework by incorporating drug-drug interaction detection into the retrieval and generation pipeline. By leveraging existing DDI databases (e.g., DrugBank, TWOSIDES), we can improve recommendations based on potential adverse drug interactions. This enables our system to prioritize safer medication choices, reducing the risk of adverse effects. Future work will also focus on improving the interpretability of the generated recommendations. Providing explanations for retrieved cases, highlighting salient features influencing predictions, and implementing attention-based visualization techniques could make the system more transparent and clinically reliable.

Table 6: F1 and Jaccard scores across different retrieval strategies. Full Sent Cosine with BioMistral achieves the best performance (F1/Jaccard = 0.52/0.34). Other strong-performing methods include Full Sent Vector and DP Pair (Visit), particularly when paired with BioMistral and Flan-T5.

Method	BioGPT		BioMistral		Phi-3		Flan-T5	
	F1	Jacc.	F1	Jacc.	F1	Jacc.	F1	Jacc.
Baseline	0.14	0.07	0.22	0.13	0.09	0.05	0.23	0.12
Full Sent. Cosine	0.22	0.13	0.52	0.34	0.12	0.07	0.29	0.17
DP Pair (Pat.)	0.16	0.12	0.36	0.24	0.02	0.01	0.29	0.17
DP Pair (Visit)	0.19	0.10	0.40	0.27	0.05	0.03	0.30	0.13
Full Sent. Vec	0.20	0.10	0.44	0.27	0.06	0.03	0.26	0.15

9 Conclusion

We introduced **TreatRAG**, a retrieval-augmented generation framework designed to enhance medication recommendation by leveraging historical patient data with pretrained language models. Evaluation on the MIMIC-IV dataset demonstrated statistically significant improvements in prediction accuracy across multiple backbone models, including BioGPT, BioMistral, Phi-3, and Flan-T5, while often simultaneously reducing the hallucination rate. TreatRAG provides an interpretable, model-agnostic approach, advancing the reliability and clinical safety of CDS systems.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. 2023. Ehrxqa: A multimodal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems* 36 (2023), 3867–3880.

[3] Tiffani J Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, et al. 2012. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine* 157, 1 (2012), 29–43.

[4] Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H Kann, Fallon E Chipidza, Jonathan Leeman, Hugo JWL Aerts, Timothy Miller, et al. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health* 6, 6 (2024), e379–e381.

[5] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua A. Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. https://api.semanticscholar.org/CorpusID:948039

[6] Yu-Neng Chuang, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2024. SPeC: a soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *Journal of Biomedical Informatics* 151 (2024), 104606.

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[8] Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, et al. 2024. TC-RAG: Turing-Complete RAG's Case study on Medical LLM Systems. *arXiv preprint arXiv:2408.09199* (2024).

[9] Alistair Johnson, Leo Bulgarelli, Tom Pollard, Brandon Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). doi:10.13026/kpb9-mt58

[10] Rainu Kaushal, Kaveh G Shojania, and David W Bates. 2003. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Archives of internal medicine* 163, 12 (2003), 1409–1416.

[11] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).

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

[13] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* 23, 6 (2022), bbac409.

[14] Rajat Mishra and S Shridevi. 2024. Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. *Scientific Reports* 14, 1 (2024), 25449.

[15] Mark A Musen, Blackford Middleton, and Robert A Greenes. 2021. Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine*. Springer, 795–840.

[16] World Health Organization. 2025. Anatomical Therapeutic Chemical (ATC) Classification. https://www.who.int/tools/atc-ddd-toolkit/atc-classification Accessed April 29, 2025.

[17] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. 314–334.

[18] Leila Shahmoradi, Reza Safdari, Hossein Ahmadi, and Maryam Zahmatkeshan. 2021. Clinical decision support systems-based interventions to improve medication outcomes: a systematic literature review on features and effects. *Medical Journal of the Islamic Republic of Iran* 35 (2021), 27.

[19] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

[20] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. GAMENet: Graph Augmented Memory Networks for Recommending Medication Combination. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Thirty-First Innovative Applications of Artificial Intelligence Conference, and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 139, 8 pages. doi:10.1609/aaai.v33i01.33011126

[21] Deepika Sharma, Gagangeet Singh Aujla, and Rohit Bajaj. 2023. RETRACTED: Evolution from ancient medication to human-centered Healthcare 4.0: A review on health care recommender systems. *International Journal of Communication Systems* 36, 12 (2023), e4058.

[22] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 2022. 4SDrug: Symptom-based Set-to-set Small and Safe Drug Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3970–3980. doi:10.1145/3534678.3539089

[23] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: adapting large language models can outperform human experts. *Research Square* (2023).

[24] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine* 6, 1 (2023), 135.

[25] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187* (2024).

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

[27] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3735–3741. doi:10.24963/ijcai.2021/514 Main Track.

[28] Hao-Ren Yao, Nairen Cao, Katina Russell, Der-Chen Chang, Ophir Frieder, and Jeremy T. Fineman. 2024. Self-Supervised Representation Learning on Electronic Health Records with Graph Kernel Infomax. *ACM Trans. Comput. Healthcare* 5, 2, Article 10 (April 2024), 28 pages. doi:10.1145/3648695

[29] Hao-Ren Yao, Der-Chen Chang, Ophir Frieder, Wendy Huang, and Tian-Shyug Lee. 2019. Multiple graph kernel fusion prediction of drug prescription. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 103–112.

[30] Hao-Ren Yao, Oskar Mencer, Han-Sun Chiang MD, Der-Chen Chang, and Ophir Frieder. 2025. Forecasting Prescription Efficacy. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 78–82. doi:10.1007/978-3-031-88720-8_14

[31] Yutao Zhang, Robert Chen, Jie Tang, Walter F. Stewart, and Jimeng Sun. 2017. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1315–1324. doi:10.1145/3097983.3098109

[32] Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. Leave No Patient Behind: Enhancing Medication Recommendation for Rare Disease Patients. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 533–542. doi:10.1145/3626772.3657785