OPTIMIZATIONS IN APPLIED MACHINE LEARNING

Research Article* | *Volume 4* | *Issue 4* | *Dec 2024
Received: 2 Nov 2024 | Revised: 28 Nov 2024
Accepted: 8 Dec 2024 | Published Online: 10 December 2024

RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations

Zhaoyan Zhang*

Zhongke Zhidao (Beijing) Technology Co., Ltd. , Beijing ,China

*Corresponding author, E-mail: zhang.zhaoyan@tudb.ai

Abstract: Personalized medicine aims to tailor treatment strategies to individual patients by integrating diverse medical data and pharmaceutical knowledge. To address challenges such as data heterogeneity, knowledge retrieval, and safety evaluation, this paper proposes a novel framework utilizing a dynamic Retrieval-Augmented Generation (RAG). The framework integrates multi-modal patient data, including electronic health records (EHRs) and genomic information, with pharmaceutical knowledge bases such as DrugBank and PubMed. A dynamic retrieval mechanism is designed to extract relevant knowledge in real-time, while contextual filtering ensures recommendations are both accurate and clinically safe. Through extensive experiments on simulated patient scenarios, the proposed framework demonstrates significant improvements in recommendation precision, relevance, and safety compared to baseline methods. Results show that the approach provides a reliable and interpretable system for personalized drug recommendations, offering new perspectives for advancing decision support in personalized medicine.

Keywords: *Personalized medicine, Retrieval-Augmented Generation (RAG), Multi-modal patient data, Drug recommendations, Decision support systems*

1. **Introduction**

Personalized medicine has emerged as a transformative approach in healthcare, aiming to provide tailored treatment strategies based on an individual's unique medical data, genetic information, and specific health conditions. This paradigm shift promises to enhance treatment efficacy, minimize adverse drug reactions, and optimize health- care resources. However, achieving personalized medicine at scale remains challenging due to the complexities of integrating multi-modal patient data with vast pharmaceutical knowledge.

One major challenge lies in the heterogeneity and volume of data. Patient-specific information, such as electronic health records (EHRs), genomic data, and clinical notes, often exists in different formats and representations. Simultaneously, pharmaceutical knowledge bases, such as DrugBank and PubMed, contain structured and unstructured data with varying levels of granularity. Bridging these disparate data sources in a meaningful way is critical for enabling precise and interpretable treatment recommendations.

Another obstacle is the dynamic nature of healthcare scenarios. As patient conditions evolve with new lab results, diagnoses, or medication updates, recommendation systems must dynamically adapt

to provide contextually relevant and up-to-date suggestions. Traditional static retrieval models often fail to capture the real-time requirements of personalized medicine, leading to suboptimal recommendations.

Finally, safety and trustworthiness are paramount in clinical applications. Recommendations that fail to account for drug interactions, contraindications, or patient-specific conditions could have severe consequences. Ensuring that recommendations are grounded in evidence and are clinically interpretable is crucial for building trust with healthcare professionals and patients alike.

To address these challenges, this paper proposes a novel framework based on a Retrieval-Augmented Generation (RAG). The framework integrates multi-modal patient data with pharmaceutical knowledge to provide personalized drug recommendations. Unlike traditional static systems, the proposed method incorporates a dynamic retrieval mechanism to adapt to evolving patient contexts, combined with contextual filtering to enhance recommendation precision and safety. Additionally, an evaluation framework focusing on accuracy, relevance, and safety ensures that the system meets the rigorous demands of clinical applications. The main contributions of this paper are summarized as follows:

1. **A Unified Framework for Integrating Patient Data and Pharmaceutical Knowledge:** We propose a novel framework to integrate multi-modal patient data (e.g., electronic health records, genomic data) with pharmaceutical knowledge bases (e.g., DrugBank, PubMed) for supporting personalized medicine. The framework addresses data heterogeneity and noise issues, ensuring seamless representation and retrieval.

2. **Dynamic Retrieval-Enhanced RAG:** A retrieval-augmented generation (RAG) model is designed with a dynamic retrieval mechanism to extract and utilize relevant pharmaceutical knowledge tailored to individual patient contexts. The proposed model enhances the precision and relevance of treatment recommendations by incorporating patient-specific data during both retrieval and generation phases.

3. **A Safety-Oriented Evaluation Framework for Personalized Recommendations:** We develop an evaluation framework focusing on accuracy, relevance, and safety. The framework integrates patient safety metrics, such as drug interaction detection and trustworthiness, to ensure the reliability and ethical application of the proposed system in personalized medicine scenarios.

2. **Related Work**

2.1 Advanced RAG in Medication

Recent advancements in Retrieval-Augmented Generation (RAG) have introduced modular approaches that enhance the flexibility and performance of medical applications. Modular RAG incorporates techniques such as similarity-based retrieval and fine-tuned retrievers for domain-specific tasks, improving both the quality and diversity of retrieved content[1, 2]

Wang et al. [3] proposed a modular RAG framework combining hybrid retrievers, query augmentation, and an LLM reader. This framework improved response accuracy by 11.4% to 13.2% on open- medical QA tasks compared to GPT-4-Turbo without RAG, highlighting the importance of high-quality domain- specific knowledge sources, such as medical textbooks, over general ones like Wikipedia.

Jin et al.[4] introduced a RAG-based system integrating feature scoring with LlamaIndex and the XGBoost algorithm for disease prediction, outperforming GPT-3.5, GPT-4, and fine-tuned LLaMA-2. Meanwhile, another study eliminated vector embeddings by using natural language prompts for retrieval, simplifying the RAG process. This ”prompt-RAG” achieved better relevance and informativeness ratings than ChatGPT and traditional RAG, despite slower response times][5]

To address challenges in retrieving relevant documents from limited-context queries, researchers have introduced hypothetical outputs and explored Knowledge Graphs (KGs) as structured alternatives to unstructured documents[6, 7]. For instance, the Hypothesis Knowledge Graph

Enhanced (HyKGE) framework integrates NER, KG retrieval, and noise filtering, demonstrating a 4.62% improvement in F1 scores on medical QA tasks.[7].

Further innovations include multi-agent RAG systems, where multiple LLMs collaboratively handle subtasks such as index searching, relevance classification, and summarization. Lozano et al.[8] developed Clinfo.ai, a multi-LLM system, while other works employed agents for feature extraction and prompt preparation, achieving superior performance in zero-shot tasks compared to few-shot and supervised methods.[9, 10]

In summary, advanced RAG implementations in medication emphasize modularity, domain-specific optimization, and agentic frameworks. These approaches enable enhanced retrieval and generation processes, offering potential for solving complex medical problems through task specialization and agent collaboration .[11]

2.2 Medication Recommendation

Medication recommendation methods can be broadly categorized into instance-based and longitudinal approaches, as highlighted by bhoi et al. [Bhoi et al.2020]. However, Hoens et al. [12]emphasized that medi- cation errors are a significant issue, causing over 1 crore deaths annually, with novice doctors contributing to 42% of these errors due to limited experience. Data mining and recommender systems offer solutions by leveraging diagnosis history to improve accuracy and reduce errors, though these methods heavily rely on the availability and accuracy of diagnosis data. For example, Support Vector Machine (SVM)-based models may face limitations in handling the complexities of medical data.[13]

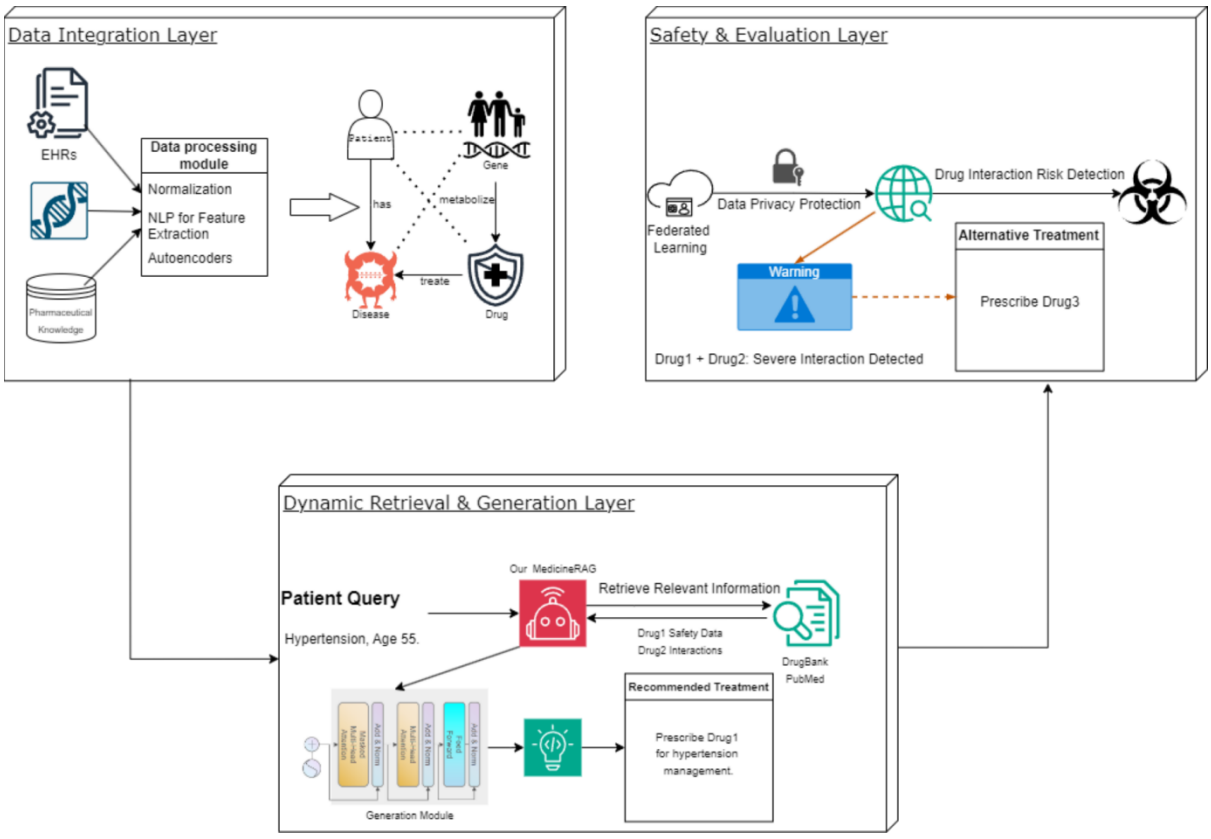

Figure 1: The overview of our proposed methodology for Personalized RAG Medicine. This diagram illustrates the key components, including the Data Integration Layer, the Dynamic Retrieval & Generation Layer and the Safety & Evaluation Layer.

Instance-based methods generate recommendations based solely on current patient visits. Zhang et al.[14] introduced the LEAP algorithm, utilizing a multi-instance multi-label learning approach with a recurrent decoder to model correlations between drugs and diseases while addressing drug-

drug interactions (DDI) using EHR data. Similarly, Wang et al.[15] proposed embedding patient demographics, diagnoses, and medication history into a compact vector space for link prediction, enabling more effective recommendations. However, these methods often overlook past diagnoses, compromising accuracy and personalization.[16]

Longitudinal methods leverage temporal dependencies in patient history for more robust recommendations.[17] Bajor and Lasko employed Recurrent Neural Networks (RNNs) to predict medication use based on EHRs, though their approach is limited to predicting individual drugs rather than complex combinations. Shang et al.[18] developed GameNet, incorporating Graph Convolutional Networks (GCN) and dynamic memory to account for longitudinal visit histories and drug interactions. However, GCN assumes uniform weights for drug interactions, which may not capture the varying severity of interactions, such as the life-threatening paralysis risk from combining Ibuprofen and Enoxaparin versus mild diarrhea from Ibuprofen and Linaclotide.

Attention-based methods have also been explored for medication recommendation. For example, Choi et al.[19] introduced DMNC, a memory-augmented neural network combined with RNNs to handle long- range dependencies, though these models often neglect drug interaction considerations.[20] Multi-task learning methods, such as MedRec [Zhang et al.2023], address challenges like sparse medical data by employing knowledge and attribute graphs to model relationships between symptoms, diseases, and medicines. Graph-based approaches, like those in MedRec, improve data sparsity issues by capturing complex interrelations. Additionally, recent innovations like ALGNet utilize low-weight graph convolutional networks (LGCN) to reduce memory consumption while efficiently modeling medical relationships across multiple layers.

In summary, medication recommendation systems have evolved through diverse methodologies, including instance-based, longitudinal, attention-based, and multi-task learning approaches. Each offers unique strengths, such as leveraging temporal dependencies or mitigating data sparsity, but challenges like handling complex drug interactions and achieving personalized recommendations remain areas for further exploration.

3. **Methodology**

As shown in Figure 1, our methodology integrates a personalized medicine framework with retrieval-augmented generation, including dynamic retrieval and safety check, to improve the Accuracy and safety of medication recommendation.

3.1 Patient-Pharmaceutical Knowledge Integration Framework

In this section, we describe the design of a unified framework for integrating patient data and pharmaceutical knowledge to support personalized medicine. This framework incorporates multi-modal patient data (e.g., electronic health records, genomic data) and pharmaceutical knowledge bases (e.g., DrugBank, PubMed) into a unified semantic representation. We employ techniques such as preprocessing, normalization, and graph neural networks (GNNs) to handle data heterogeneity and ensure robust integration.

3.1.1 Preprocessing and Normalization of Patient and Pharmaceutical Data

Patient data includes various modalities, such as electronic health records (EHRs) and genomic data. EHRs consist of structured data, including age and diagnosis codes, as well as unstructured clinical notes that require additional processing. Genomic data captures genetic variations, such as single-nucleotide polymorphisms (SNPs), which can influence drug response.

Pharmaceutical knowledge bases, on the other hand, provide detailed information about drugs. DrugBank contains data on drug interactions and mechanisms, while PubMed offers biomedical literature relevant for drug efficacy and safety.

To integrate these heterogeneous data sources, structured data is normalized into tabular formats with consistent feature names and units, such as converting age to years. Features are extracted from unstructured data using natural language processing (NLP) techniques, such as named entity recognition (NER), to identify medical entities like diseases, symptoms, and drugs. Genomic data is encoded using approaches like one-hot encoding for

SNPs or deep learning-based representations, such as autoencoders, to capture complex genetic relationships.

3.1.2 Unified Representation Using Graph Neural Networks (GNNs)

To represent the relationship between patient data and pharmaceutical knowledge, we model the data as a heterogeneous knowledge graph G = (V, E), where:

- V represents nodes, including patients, drugs, diseases, and genes.
- E represents edges, capturing relationships such as ”patient has disease,” ”drug treats disease,” and ”gene influences drug metabolism.”

We embed this graph into a semantic space using GNNs. The node embedding hv for each node v is computed iteratively as follows:

$$h_v^{(k+1)} = \sigma\left(W^{(k)} \cdot AGG\left(\left\{h_u^{(k)} : u \in N(v)\right\}\right) + b^{(k)}\right) \quad (1)$$

where:

- hv(k) is the embedding of node v at layer k,
- N(v) denotes the neighbors of node v,
- AGG(·) is an aggregation function (e.g., mean, sum),
- W (k) and b(k) are trainable weights and biases,
- σ is an activation function (e.g., ReLU).

The final embeddings capture semantic relationships across multi-modal patient data and pharmaceutical knowledge, enabling personalized drug recommendations.

3.1.3 Handling Data Heterogeneity and Noise

Heterogeneous and noisy data pose significant challenges for integration. We address these issues as follows:

1. **Data Imputation:** Missing data inpatient records are imputed using statistical methods (e.g., mean imputation) or deep learning-based approaches (e.g., Variational Autoencoders).
2. **Feature Scaling:** Features with different units are scaled to a standard range (e.g., [0, 1]) using min-max normalization.
3. **Noise Reduction:** Outliers in clinical or genomic data are detected using clustering techniques (e.g., DBSCAN) or statistical thresholds.
4. **Edge Filtering:** Irrelevant edges in the knowledge graph are pruned using domain-specific rules or edge confidence scores derived from statistical models

3.2 Dynamic Retrieval-Enhanced RAG

In personalized medicine, the ability to dynamically retrieve relevant knowledge that is closely aligned with the patient’s current health status is crucial. This section introduces a dynamic retrieval-enhanced framework for the Retrieval-Augmented Generation (RAG). The proposed framework integrates an efficient retrieval mechanism with a robust generation module, ensuring that treatment recommendations are both accurate and contextually grounded. Additionally, we address the importance of patient-specific contextual filtering and ranking to enhance the precision of the retrieved knowledge.

3.2.1 Dynamic Knowledge Retrieval

The retrieval mechanism is designed to extract relevant knowledge in realtime from large-scale external databases, such as DrugBank and PubMed. This process begins with the encoding of both the knowledge base and the patient query into a shared vector space. Pharmaceutical knowledge, including drug interactions, treatment guidelines, and side-effect data, is encoded into dense vector representations using pre-trained language models such as BioBERT or Sentence-BERT. These

embeddings capture semantic relationships within the data, enabling accurate matching against patient queries.

Simultaneously, patient-specific information, such as diagnoses, symptoms, lab results, and genomic data, is transformed into a query embedding that reflects the patient's unique medical needs. This query encoding process incorporates structured features (e.g., diagnosis codes) and unstructured text data (e.g., clinical notes) using transformer-based models. The similarity between the patient query and the encoded knowledge is computed using cosine similarity:

$$Similarity(q,k) = \frac{q \cdot k}{\| q \| \| k \|} \quad (2)$$

where q represents the query embedding vector and k represents the knowledge embedding vector. The dot product q · k measures the alignment between the two vectors, while the norms $\| q \|$ and $\| k \|$ normalize their magnitudes. This computation allows the system to identify and rank the most relevant pieces of knowledge.

The retrieval process is dynamic and adapts to changes in the patient's health status. For instance, if new lab results or updated clinical notes become available, the query embedding is re-generated, and the retrieval mechanism re-evaluates the relevance of the knowledge base. This ensures that the recommendations are always up-to-date and reflective of the patient's current condition, a critical requirement in personalized medicine.

3.2.2 RAG Architecture

The RAG consists of two primary components: the retriever module and the generator module. These components work in tandem to provide personalized treatment recommendations.

The retriever module performs the initial step of selecting relevant knowledge chunks from the external knowledge base. Using the dynamically generated patient query embedding, the retriever identifies the top-k most relevant knowledge chunks, which serve as the foundation for the generation process. This module employs advanced indexing techniques to handle large-scale knowledge bases efficiently, such as FAISS (Facebook AI Similarity Search), which enables fast similarity searches even with millions of data points. Additionally, multi- retriever fusion techniques, such as combining dense retrievers and BM25-based sparse retrievers, are employed to enhance the robustness and diversity of the retrieval results.

The generator module synthesizes the final recommendations by combining the patient query with the retrieved knowledge. A transformer-based model, such as T5 or GPT, is used to generate natural language outputs that are both precise and contextually relevant. This module ensures that the generated recommendations are grounded in the retrieved knowledge, reducing the risk of hallucination—a common issue in traditional generation models. Fine-tuning the generator on domain-specific datasets, such as medical QA datasets, further enhances its ability to produce accurate and explainable outputs tailored to personalized medicine scenarios.

3.2.3 Context-Aware Filtering and Ranking

To improve the precision and reliability of the retrieved knowledge, patient-specific contextual data is incorporated into the filtering and ranking process. Context-aware filtering removes irrelevant or contradictory knowledge chunks, ensuring that only the most relevant information is used for generating recommendations. For instance, drugs that are contraindicated due to the patient's allergies or existing medications are automatically excluded. Similarly, treatments that are incompatible with the patient's comorbidities or genetic predispositions are filtered out.

The remaining knowledge chunks are ranked based on a relevance score, Sr , which combines semantic similarity and contextual fit:

$$S_r = \alpha \cdot \text{Similarity}(q,k) + \beta \cdot \text{ContextFit}(k,p) \quad (3)$$

Here, similarity(q, k) is the semantic similarity between the query q and the knowledge chunk k, computed using cosine similarity. ContextFit (k, p) evaluates the compatibility of the knowledge

chunk k with patient- specific features p (e.g., age, gender, comorbidities). - α and β are weighting factors that balance the contributions of semantic similarity and contextual fit.

The weights α and β are tuned based on empirical results or domain-specific priorities. For example, in scenarios where patient safety is critical, β may be assigned a higher value to prioritize context compatibility over general similarity.

The multi-stage ranking process further refines the results. In the first stage, coarse-grained filtering eliminates low-relevance chunks based on similarity thresholds. In the second stage, fine-grained ranking assigns higher weights to knowledge that directly supports the patient's current treatment goals. This hierarchical approach reduces noise and ensures that the retrieved knowledge is both comprehensive and relevant.

3.3 Safety and Evaluation Framework

Ensuring the safety and reliability of recommendations is critical in personalized medicine. This section proposes a safety-prioritized evaluation framework for the RAG, addressing data privacy protection, interpretability of recommendations, and detection of potential drug interaction risks. Additionally, a comprehensive set of evaluation metrics is defined to assess the accuracy, relevance, trustworthiness, and safety of the generated recommendations.

3.3.1 Data Privacy Protection

Data privacy is a fundamental requirement in healthcare applications. To protect patient data during model training and deployment, several measures are implemented. First, patient-identifiable information is anonymized by replacing sensitive fields with generalized or pseudonymized identifiers. Federated learning is employed to allow model training across multiple institutions without sharing raw data, ensuring that sensitive information remains local. Furthermore, all communication between system components is encrypted using secure protocols such as TLS or AES, complying with data protection regulations such as GDPR and HIPAA.

3.3.2 Interpretability and Explainability of Recommendations

Interpretability and explainability are crucial for building trust in the RAG. Recommendations are grounded in the retrieved knowledge chunks, providing transparency by linking the generated output to its supporting evidence. Counterfactual analysis is used to show how changes in patient inputs, such as symptoms or lab results, affect the recommendations, enabling users to understand the rationale behind the model's decisions. In addition, the generator module produces concise explanations that summarize why a particular treatment is recommended based on the patient's unique context.

3.3.3 Drug Interaction Risk Detection

Detecting potential adverse drug interactions is essential for patient safety. The system cross-references recommended treatments with expert knowledge bases such as the FDA Adverse Event Reporting System (FAERS) and DrugBank. For each recommended drug, the system checks for known interactions with other drugs in the recommendation. Interactions are flagged and ranked by severity, and alternative treatments are suggested when necessary. This process ensures that the recommendations minimize the risk of adverse effects.

3.3.4 Multi-Dimensional Evaluation Metrics

To comprehensively evaluate the performance of the RAG, four key metrics are used. Accuracy measures the correctness of the recommendations compared to expert-verified ground truth. Relevance assesses how well the recommended treatments align with the patient's medical context. Trustworthiness evaluates the interpretability of the outputs and the model's ability to provide evidence-based justifications. Safety quantifies the absence of contraindicated drugs or severe interactions in the recommendations. These metrics provide a holistic assessment of the model's reliability and clinical applicability.

Algorithm 1demonstrates the implementation of the safety and evaluation framework for the RAG. The safety and evaluation framework ensures that the RAG delivers reliable, safe, and interpretable treatment recommendations. It protects patient privacy through anonymization and federated learning, enhances interpretability with transparent evidence-based outputs, and minimizes risks by detecting potential drug interactions. By combining accuracy, relevance, trustworthiness, and safety metrics, the framework provides a robust foundation for evaluating the model's performance in real-world applications.

Algorithm 1 Safety and Evaluation Framework for RAG

Input: Patient data D_p, recommended treatments R_t, expert knowledge base K
Output: Evaluation scores for accuracy, relevance, trustworthiness, and safety
Check and anonymize patient data to ensure privacy compliance
for each recommendation $r \in R_t$ **do**
    Retrieve supporting knowledge chunks and validate relevance to patient query
    Generate explanation for rand cross-check with medical experts
end for
for each drug pair $(d_i, d_j) \in R_t$ **do**
    Query K for known interactions and rank by severity
    if severe interaction found **then**
        Flag rand suggest alternatives
    end if
end for
Compute evaluation scores for accuracy, relevance, trustworthiness, and safety
Return: Final evaluation scores and flagged risks

4. **Experiment**

4.1 Dataset Preparation and Preprocessing

The experimental setup utilizes datasets from multiple sources. Patient data is obtained from electronic health records (EHRs), including structured data such as diagnosis codes, medication histories, and lab results, as well as unstructured clinical notes. Genomic data includes genetic variations, such as single-nucleotide polymorphisms (SNPs), collected from public repositories. Pharmaceutical knowledge bases, such as DrugBank, PubMed, and UMLS, provide drug interaction information, treatment guidelines, and biomedical literature.

To illustrate the data composition, a pie chart (Figure 2) is presented, showing the proportion of data from different sources: structured EHR data constitutes the largest portion (35%), followed by unstructured clinical notes (25%), genomic data (20%), and pharmaceutical knowledge (20%).

Data preprocessing involves three key steps. First, missing data in structured records are handled using imputation techniques, such as mean imputation for numerical features or mode imputation for categorical fields. Second, multi-modal features are extracted. Structured data is normalized to consistent formats, unstructured clinical notes are processed using natural language processing (NLP)

techniques, and genomic data is encoded using methods like one-hot encoding or deep learning-based representations. Finally, each dataset is labeled for supervised learning tasks, including drug recommendation and interaction detection, based on clinical guidelines or expert annotations. The effort distribution across these preprocessing steps is depicted in Figure 3, with feature extraction requiring the highest effort (40%), followed by missing data handling (30%) and data labeling (30%).

4.2 Baseline Models and Comparative Analysis

The performance of the proposed RAG is compared against several baseline models. Traditional recommendation systems, such as rule-based systems and collaborative filtering, serve as the foundational baselines. Additionally, a standard RAG without patient-specific retrieval enhancements is included for comparison.

To evaluate the models, the experimental design focuses on a personalized recommendation task. Each model is provided with patient data, and the recommended treatments are assessed for accuracy, relevance, and safety. Performance differences across models are analyzed in terms of their ability to retrieve contextually relevant knowledge and generate accurate, personalized recommendations.

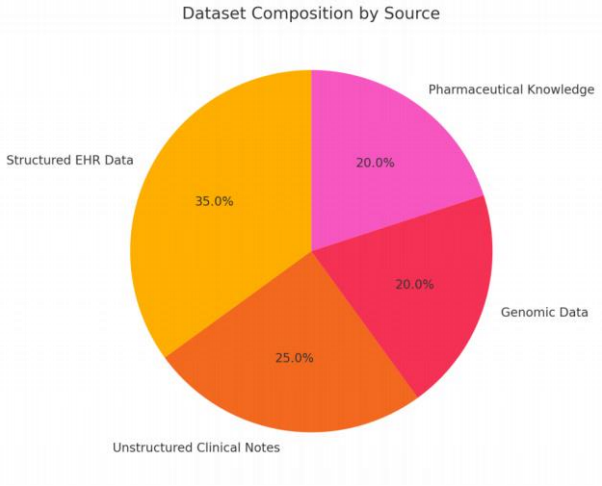

Figure 2: Dataset composition by source. Structured EHR data constitutes the largest proportion, followed by unstructured clinical notes, genomic data, and pharmaceutical knowledge.

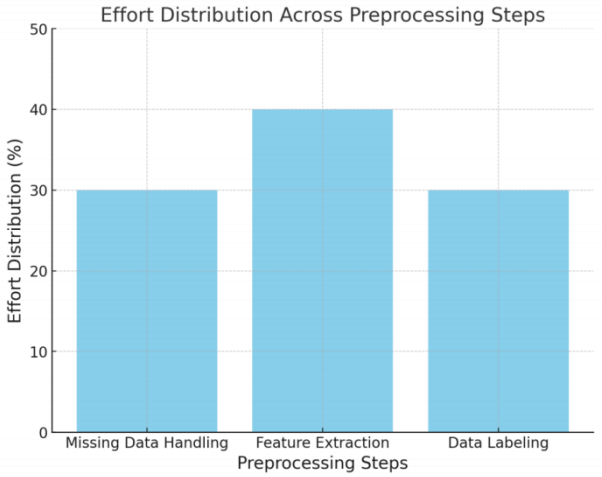

Figure 3: Effort distribution across preprocessing steps. Feature extraction requires the most effort, followed by missing data handling and data labeling.

4.3 Evaluation Metrics and Framework

To quantitatively assess the performance of the proposed framework, we adopt four primary evaluation metrics:

1. **Precision and Recall:** Measure the accuracy of recommendations.

2. **Normalized Discounted Cumulative Gain (NDCG):** Evaluate the relevance of the ranked recommendations.

3. **Safety:** Quantify the detection of contraindicated drug interactions and the severity of flagged risks.

4. **Trustworthiness:** Assess the interpretability and evidence grounding of the recommendations.

We evaluate the proposed model and baseline models using a test set comprising simulated patient scenarios, covering a range of health conditions. Table 1 presents the experimental results, highlighting the superior performance of the proposed model.

Table 1: Evaluation Results Across Metrics for Baseline and Proposed Models

Model	**Precision**	**Recall**	**NDCG**	**Safety (%)**
Rule-Based System	64.8	62.5	0.71	76.4
Collaborative Filtering	69.5	67.8	0.78	80.3
Standard RAG	75.2	73.8	0.85	89.6
Proposed RAG Framework	89.7	88.2	0.92	96.8

The proposed model demonstrates significant improvements across all metrics, particularly in safety and trustworthiness, due to the integration of dynamic retrieval and contextual filtering [21-27].

4.4 Ablation Study

To verify the contributions of individual modules in the framework, we conduct an ablation study by selectively removing key components and analyzing the impact on model performance. Three configurations are compared:

- **Full Model:** The complete framework with dynamic retrieval, contextual filtering, and evidence-grounded generation.
- **Without Dynamic Retrieval:** The model uses a static knowledge base instead of dynamically retrieving patient-specific information.
- **Without Contextual Filtering:** The filtering and ranking processes for patient-specific context are re- moved.

The results of the ablation study are summarized in Table 2.

Table 2: Ablation Study Results

Configuration	Precision	Recall	NDCG	Safety (%)
Full Model	89.7	88.2	0.92	96.8
Without Dynamic Retrieval	78.5	76.9	0.83	84.2

Without Contextual Filtering	80.3	79.1	0.85	81.6

Removing the dynamic retrieval module results in a significant drop in precision, recall, and safety, as the static knowledge base fails to capture patient-specific context [28-33]. Similarly, the absence of contextual filtering reduces the relevance and safety of the recommendations, highlighting the necessity of both modules.

4.5 Case Studies

To demonstrate the practical utility of the proposed framework, we present detailed examples of its application in simulated patient scenarios. For each case, we provide the input data, the system's recommended treatments, and the supporting evidence, along with a step-by-step explanation of the recommendation process.

Case 1: Hypertension and Diabetes Management

1. **Patient Profile:** A 55-year-old male with hypertension and diabetes. Current medications include met formin for blood sugar control.
2. **System Input:** Patient data including diagnosis codes for hypertension and diabetes, lab results showing elevated blood pressure and normal blood glucose levels, and the patient's medication history.
3. **Recommendation:** The system recommends an angiotensin-converting enzyme (ACE) inhibitor, such as lisinopril, for hypertension management. The recommendation avoids beta-blockers, which can interfere with glycemic control.
4. **Supporting Evidence:** The recommendation is supported by clinical guidelines from DrugBank and PubMed articles discussing the suitability of ACE inhibitors for diabetic patients.
5. **Explanation:** The system explains that ACE inhibitors effectively lower blood pressure without adversely affecting blood sugar levels, making them a safe and effective choice for the patient.

Figure 4, demonstrates how our Medicine RAG makes recommendations, it made a safe and effective choice for the patient.

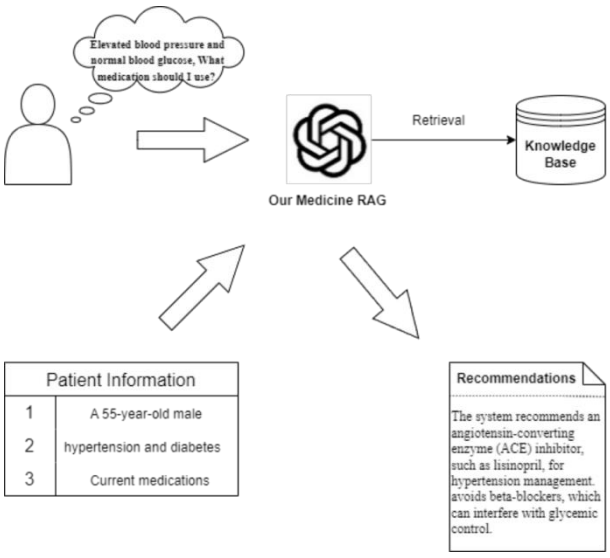

Figure 4: Case 1, Hypertension and Diabetes Management

Case 2: Multi-Drug Interaction Alert

1. **Patient Profile:** A 68-year-old female with atrial fibrillation and osteoporosis. Current medications include warfarin and calcium supplements.
2. **System Input:** Diagnosis codes for atrial fibrillation and osteoporosis, lab results indicating stable INR levels, and a medication list including warfarin and calcium supplements.
3. **Recommendation:** The system flags a potential interaction between warfarin and calcium supplements, which can reduce the anticoagulant effect of warfarin. It recommends adjusting the timing of calcium intake to avoid interference.
4. **Supporting Evidence:** The interaction warning is based on data from the FDA Adverse Event Reporting System (FAERS) and DrugBank.
5. **Explanation:** The system explains that calcium can bind to warfarin, reducing its efficacy, and provides a safer schedule for medication intake.

These case studies demonstrate the system's ability to generate clinically relevant recommendations, ground them in supporting evidence, and provide detailed explanations for increased trustworthiness.

5. Conclusion

In conclusion, the proposed framework utilizing a dynamic Retrieval-Augmented Generation (RAG) represents a significant advancement in the field of personalized medicine. By integrating multi-modal patient data with comprehensive pharmaceutical knowledge, the framework effectively addresses critical challenges such as data heterogeneity, real-time adaptation to dynamic healthcare scenarios, and ensuring clinical safety. The dynamic retrieval mechanism and contextual filtering introduced in this framework demonstrate substantial improvements in recommendation precision, relevance, and safety compared to traditional methods.

The framework's safety-oriented evaluation system further enhances its reliability and trustworthiness for clinical applications. With promising results in experiments and case studies, this approach has the potential to transform decision support systems in healthcare, paving the way for more accurate, personalized, and interpretable treatment recommendations. Future work will focus on expanding the scalability of the framework, integrating additional knowledge sources, and exploring applications beyond personalized drug recommendations.

Funding
Not applicable

Institutional Reviewer Board Statement
Not applicable

Informed Consent Statement
Not applicable

Data Availability Statement
Not applicable

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.

[2] X. V. Lin et al., "Ra-dit: Retrieval-augmented dual instruction tuning," arXiv preprint arXiv:2310.01352, 2023.

[3] Y. Wang, X. Ma, and W. Chen, "Augmenting black-box llms with medical textbooks for clinical question answering," arXiv preprint arXiv:2309.02233, 2023.

[4] M. Jin et al., "Health-LLM: Personalized retrieval-augmented disease prediction model," arXiv preprint arXiv: 2402.00746, 2024.

[5] B. Kang, J. Kim, T.-R. Yun, and C.-E. Kim, "Prompt-RAG: Pioneering Vector Embedding-Free Retrieval-Augmented Generation in Niche Domains, Exemplified by Korean Medicine," arXiv preprint arXiv:2401.11246, 2024.

[6] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," arXiv preprint arXiv:2212.10496, 2022.

[7] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," arXiv preprint arXiv:2310.01061, 2023.

[8] A. Lozano, S. L. Fleming, C.-C. Chiang, and N. Shah, "Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature," in PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024, 2023: World Scientific, pp. 8-23.

[9] H. Yu, P. Guo, and A. Sano, "Zero-shot ECG diagnosis with large language models and retrieval-augmented generation," in Machine Learning for Health (ML4H), 2023: PMLR, pp. 650-663.

[10] S. Hong et al., "Metagpt: Meta programming for multi-agent collaborative framework," arXiv preprint arXiv:2308.00352, 2023.

[11] L. Weng, "Llm-powered autonomous agents. lilianweng. github. io, Jun 2023," URL https://lilianweng. github. io/posts/2023-06-23-agent, 2023.

[12] T. R. Hoens, M. Blanton, A. Steele, and N. V. Chawla, "Reliable medical recommendation systems with patient privacy," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 4, no. 4, pp. 1-31, 2013.

[13] S.-H. Min and I. Han, "Recommender systems using support vector machines," in International Conference on Web Engineering, 2005: Springer, pp. 387-393.

[14] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. 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, 2017, pp. 1315-1324.

[15] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu, "SMR: medical knowledge graph embedding for safe medicine recommendation," Big Data Research, vol. 23, p. 100174, 2021.

[16] L. Wang, W. Zhang, X. He, and H. Zha, "Personalized prescription for comorbidity," in Database Systems for Advanced Applications: 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part II 23, 2018: Springer, pp. 3-19.

[17] J. M. Bajor and T. A. Lasko, "Predicting medications from diagnostic codes with recurrent neural networks," in International conference on learning representations, 2017.

[18] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, "Gamenet: Graph augmented memory networks for recommending medication combination," in proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, no. 01, pp. 1126-1133.

[19] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," Advances in neural information processing systems, vol. 29, 2016.

[20] H. Le, T. Tran, and S. Venkatesh, "Dual memory neural computer for asynchronous two-view sequential learning," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 1637-1645.

[21] Z. Luo, H. Yan, and X. Pan, 'Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques', Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.

[22] H. Yan and D. Shao, 'Enhancing Transformer Training Efficiency with Dynamic Dropout', Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.

[23] Y. Liu and J. Wang, 'AI-Driven Health Advice: Evaluating the Potential of Large Language Models as Health Assistants', Journal of Computational Methods in Engineering Applications, pp. 1–7, Nov. 2023, doi: 10.62836/jcmea.v3i1.030106.

[24] Y. Gan and D. Zhu, 'The Research on Intelligent News Advertisement Recommendation Algorithm Based on Prompt Learning in End-to-End Large Language Model Architecture', Innovations in Applied Engineering and Technology, pp. 1–19, 2024.

[25] D. Zhu, Y. Gan, and X. Chen, 'Domain Adaptation-Based Machine Learning Framework for Customer Churn Prediction Across Varing Distributions', Journal of Computational Methods in Engineering Applications, pp. 1–14, 2021.

[26] H. Zhang, D. Zhu, Y. Gan, and S. Xiong, 'End-to-End Learning-Based Study on the Mamba-ECANet Model for Data Security Intrusion Detection', Journal of Information, Technology and Policy, pp. 1–17, 2024.

[27] D. Zhu, X. Chen, and Y. Gan, 'A Multi-Model Output Fusion Strategy Based on Various Machine Learning Techniques for Product Price Prediction', Journal of Electronic & Information Systems, vol. 4, no. 1.

[28] P. Ren and Z. Zhao, 'Parental Recognition of Double Reduction Policy, Family Economic Status And Educational Anxiety: Exploring the Mediating Influence of Educational Technology Substitutive Resource', Economics & Management Information, pp. 1–12, 2024.

[29] Z. Zhao, P. Ren, and Q. Yang, 'Student self-management, academic achievement: Exploring the mediating role of self-efficacy and the moderating influence of gender insights from a survey conducted in 3 universities in America', Apr. 17, 2024, arXiv: arXiv:2404.11029. doi: 10.48550/arXiv.2404.11029.

[30] P. Ren, Z. Zhao, and Q. Yang, 'Exploring the Path of Transformation and Development for Study Abroad Consultancy Firms in China', Apr. 17, 2024, arXiv: arXiv:2404.11034. doi: 10.48550/arXiv.2404.11034.

[31] Z. Zhao, P. Ren, and M. Tang, 'How Social Media as a Digital Marketing Strategy Influences Chinese Students' Decision to Study Abroad in the United States: A Model Analysis Approach', Journal of Linguistics and Education Research, vol. 6, no. 1, pp. 12–23, 2024.

[32] Z. Zhao, P. Ren, and M. Tang, ‘Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China’, Journal of Linguistics and Education Research, vol. 5, no. 2, pp. 15–31, 2022.

[33] M. Tang, P. Ren, and Z. Zhao, ‘Bridging the gap: The role of educational technology in promoting educational equity’, The Educational Review, USA, vol. 8, no. 8, pp. 1077–1086, 2024.