Experience Retrieval-Augmentation with Electronic Health Records Enables Accurate Discharge QA

Justice Ou[1 * †], **Tinglin Huang**[2 †], **Yilun Zhao**[2], **Ziyang Yu**[3], **Yifei Shen**[4], **Peiqing Lu**[5], **Rex Ying**[2]
University of Illinois Urbana-Champaign[1], Yale University[2],
University of Waterloo[3], University of Washington[4], Microsoft[5]

Abstract

To improve the reliability of Large Language Models (LLMs) in clinical applications, retrieval-augmented generation (RAG) is extensively applied to provide factual medical knowledge. Beyond general medical knowledge, clinical case-based knowledge is also critical for effective medical reasoning, as it provides context grounded in real-world patient experiences. Motivated by this, we propose Experience Retrieval-Augmentation (EXPRAG) framework based on Electronic Health Record (EHR), aiming to offer the relevant context from other patients' discharge reports. EXPRAG performs retrieval through a coarse-to-fine process: it first applies an EHR-based report ranker to efficiently identify similar patients as experience, and then utilizes a context retriever to extract task-relevant content for enhanced medical reasoning. To evaluate RAG systems on EHR data including EXPRAG and medical agents, we introduce DISCHARGEQA, a clinical QA dataset with 1,280 discharge-related questions across diagnosis, medication, and instruction tasks. Each problem is generated using historical EHR data to ensure realistic and challenging scenarios. Experimental results demonstrate that EXPRAG consistently outperforms traditional text-based rankers, achieving an average relative improvement of 5.2%, highlighting the importance of case-based knowledge for medical reasoning.

1 Introduction

Benefiting from pretraining on large-scale corpora, Large Language Models (LLMs) are capable of performing complex reasoning and have shown great promise in medical applications (Zheng et al., 2024; Liu et al., 2024). One important application is inferring clinical conditions, including diagnosis and medication, which can be formulated as a question-answering (QA) task (Singhal et al., 2025; Chen et al., 2023; Huang et al., 2024). However, LLM agents often suffer from hallucinations and a lack of domain-specific knowledge, which limits their reliability in real-world medical applications.

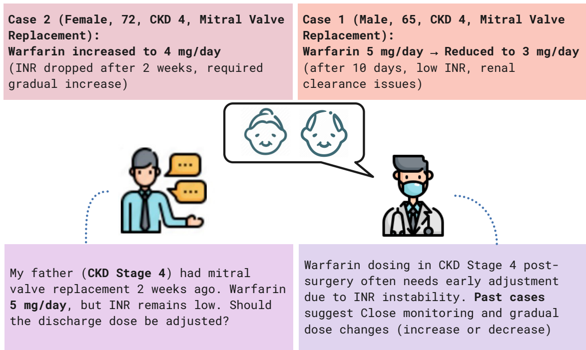

Figure 1: An illustrative example of utilizing experience from relevant clinical cases to support medical decision: adjusting a patient's warfarin dosage based on the specific clinical context rather than relying on a generic standard dose.

To address this, prior studies have resorted to retrieving factual knowledge from open-ended databases to provide context, such as the description of drugs from Wikipedia (Xiong et al., 2024a; Yang et al., 2025). Such external knowledge enables LLMs to access general medical facts, thereby improving response accuracy. However, introducing such general facts cannot effectively help LLMs solve real clinical cases, which often involve coexisting clinical conditions. For example, as shown in Figure 1, adjusting a patient's warfarin dosage requires reasoning based on the specific clinical context, whereas conventional retrieval can only provide the standard dosage for warfarin, which is irrelevant in this case.

In light of this, we argue that, in addition to general factual concepts, clinical case-based knowl-

*Correspondence to zo6@illinois.edu
†The two first authors made equal contributions.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23406–23430

edge is also crucial for effective medical reasoning. The intuition is that an experienced clinician often relies on past cases with similar conditions to guide diagnosis, treatment decisions, and discharge planning. To this end, we propose **Experience Retrieval-Augmentation (ExpRAG)** framework, leveraging a large-scale EHR database MIMIC-IV (Johnson et al., 2023b) as its knowledge basis. Specifically, ExpRAG breaks down the retrieval process into two steps: (1) report ranking applies an EHR-based similarity measurement to identify patients with similar medical conditions, and (2) experience retrieval extracts problem-relevant content from these patients' discharge reports, which serves as the case-based contextual knowledge for LLMs. The introduced EHR modality enables large-scale clinical experience retrieval, grounding the model's reasoning in real-world clinical practices.

To evaluate capability of ExpRAG and other RAG methods/agents in medical reasoning, we introduce DischargeQA, a clinical dataset including 1,280 QA pairs dedicated to discharge-related problems. The dataset primarily includes three types of problems:simulating the discharge process of final diagnosis, medication prescription and post-discharge instructions. For each problem, we follow the data structure of the discharge report and select the content preceding the question as the problem background to avoid label leakage. Additionally, we index the option candidates using EHR to generate contextually relevant options, ensuring a non-trivial and clinically meaningful challenge for the model.

We evaluate the performance of five different LLMs and compare ExpRAG with the text-based report ranker using DischargeQA. [1]

The results demonstrate the effectiveness of using EHR to retrieve relevant clinical experience, as it consistently improves the performance of LLM backbones and outperforms the text-based ranker with an average relative improvement of 5.2%. Our main contributions are summarized as follows:

- We propose ExpRAG, an EHR-based experience retrieval-augmentation framework, shedding light on the potential of leveraging past clinical cases to enhance LLM performance in medical reasoning tasks.
- We introduce DischargeQA, a medical QA dataset for discharge-related questions, designed to evaluate LLMs' ability to simulate the clinical decision-making process during patient discharge with a more challenging setup.
- Our results demonstrate the advantage of ExpRAG over the text-based ranker, highlighting the effectiveness of EHR in providing clinically meaningful context.

[1] DischargeQA on https://physionet.org/ , please check updates on Github: https://github.com/jou2024/EXPRAG

2 Related Work

Retrieval-Augmented Generation (RAG). RAG has become a key paradigm for overcoming the static knowledge limitations of LLMs by retrieving external information (Gu et al., 2018; Petroni et al., 2019). Traditional RAG frameworks typically use dense retrieval methods to augment generative tasks (Devlin et al., 2019; Xiong et al., 2021). While effective in general QA, these methods often lack domain specificity, which is critical in healthcare (Lu et al., 2024). Recent advancements like ClinicalRAG (Lu et al., 2024) MIRAGE (Xiong et al., 2024a) and i-MedRAG (Xiong et al., 2024b) address this by integrating structured EHR data and clinical notes for diagnosis and treatment planning. Latest work like MedRAG (Zhao et al., 2025) KARE (Jiang et al., 2025) build Knowledge Graph, RGAR (Liang et al., 2025) focuses on recurrent fusion across knowledge sources, none exploit case-level similarity inside large EHR cohorts. However, existing benchmarks primarily focus on isolated information retrieval (Johnson et al., 2023a; Kweon et al., 2024), overlooking the complexities of reasoning over patient histories and similar cases. To bridge this gap, our work extends RAG by combining structured EHR data with discharge summaries, enabling experience-driven reasoning for more realistic and reliable medical QA.

Medical QA Benchmark. EHRSQL (Lee et al., 2022) CLINSQL (Shen et al., 2026) and DrugEHRQA (Bardhan et al., 2022) target structured data queries, with the formers addressing SQL-based operations and the latter focusing on drug-related questions. EHRNoteQA (Kweon et al., 2024) and RadQA (Soni et al., 2022) leverage clinician-verified QA pairs from discharge summaries and radiology reports, while MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019) evaluate LLMs with questions from medical exams or PubMed articles. Discharge-summary-focused datasets like em-

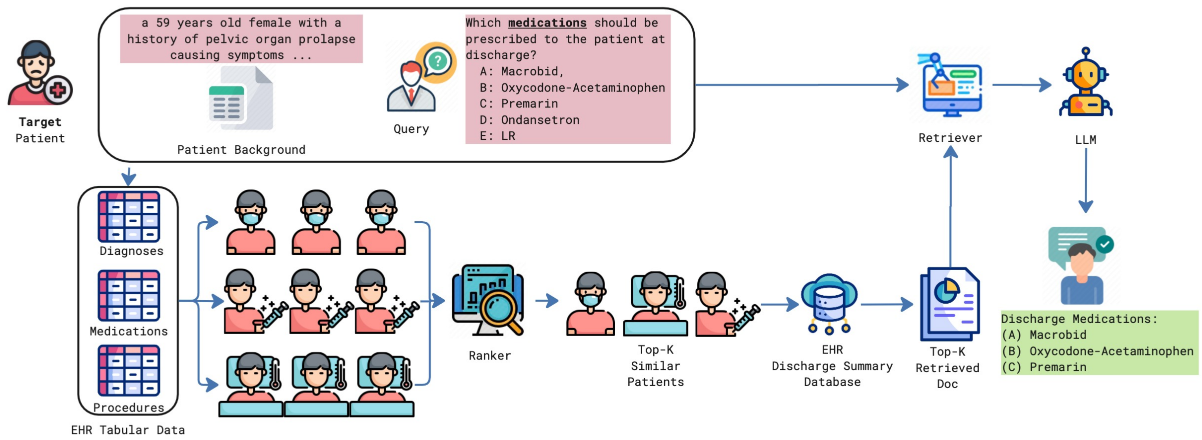

Figure 2: The overview of EXPRAG: Given a medical query and the patient's background, EXPRAG first indexes similar patients based on diagnosis, medication, and procedure similarity from the EHR. A text retriever is then applied to the discharge reports of the top-ranked similar patients to extract clinically relevant content, which is subsequently fed into the LLM to generate the answer.

rQA (Pampari et al., 2018) and CliniQG4QA (Yue et al., 2021) use discharge notes for QA tasks, and specialized datasets like RxWhyQA (Fan, 2019) and drug-reasoning QA (Moon et al., 2023) focus on specific question types like medication reasoning. PMC-Patients (Zhao et al., 2023) further studies retrieval for clinical decision support using literature-derived patient summaries, including patient-to-patient retrieval. In contrast, DISCHARGEQA is built directly from hospital EHR and discharge summaries, and evaluates downstream discharge decision-making rather than retrieval quality alone. Specificall, DISCHARGEQA introduces an evaluation framework centered around the discharge process, simulating the clinical workflow from diagnosis inference to medication prescription and discharge instruction generation. Additionally, we leverage EHR to generate non-trivial, contextually relevant candidate options, providing a more challenging and realistic setup.

3 EXPRAG Framework

EXPRAG provides a comprehensive framework for retrieving relevant knowledge from the cohort, as shown in Figure 2. In this section, we first formulate the problem that EXPRAG aims to tackle and then elaborate the two-step retrieval framework.

3.1 Task Formulation

A cohort contains a set of discharge report $\mathcal{D} = \{D_i\}_N$ where $D_i = \{d_j\}_M$ denotes the i-th report and d_j is j-th paragraph in D_i. Each report is a medical document that offers an overview of a patient's hospitalization. The goal of EXPRAG is to extract relevant content from $\mathcal{D}$ that helps LLM to effectively answer a given medical query q related to a specific patient p:

$$d_* = f_{\text{EXPRAG}}(p, q, \mathcal{D}) \quad (1)$$

The queries studied in this work focus on providing professional medical guidance for patients, including diagnosis, medication, and discharge instructions, thereby simulating realistic and practical clinical scenarios, as discussed in Section 4.

Different from the conventional RAG focusing on extracting factual concepts from open-ended databases, EXPRAG aims to utilize contextually-relevant clinical practice, inspired by how doctors collect and apply experience from past clinical cases. These two approaches rely on different reasoning procedures and knowledge sources, making them complementary to each other.

3.2 Coarse-to-Fine Retrieval Framework

To retrieve information from the cohort, one naive solution is to concatenate all the reports into one document and apply a text retriever to extract relevant content, similar to the conventional RAG pipeline. However, a standard EHR cohort typically contains millions of hospital visits, making it impractical to exhaustively search over all reports.

To efficiently perform experience retrieval, EXPRAG applies a two-step framework which conduct the retrieval from a coarse to fine level:

Report Ranking. Before addressing a specific medical query, an intuitive assumption is that only patients with similar clinical histories, e.g., similar diseases or medications, can potentially provide meaningful guidance. In light of this, EXPRAG first employs a report ranker to efficiently discard unrelated cases and narrow down the candidate pool using the patient information:

$$\mathcal{D}' = f_{\text{Ranker}}(p, \mathcal{D}) \tag{2}$$

where $\mathcal{D}' = \{D_i\}_{N' \ll N}$ is a small subset of the selected discharge summaries. The ranker module will be scalable and enable effective utilization of patient context. Specifically, we introduce EHR as a knowledge base to facilitate the patient-level similarity measurement, as presented in Section 3.3.

Experience Retrieval. Based on the selected candidate pool, a traditional text retriever is capable of providing more accurate and dedicated clinical experience searching:

$$d_* = f_{\text{Retriever}}(q, \mathcal{D}') \tag{3}$$

Built on top of clinically relevant reports identified by the dedicated ranking approach, the retriever focuses on extracting content related to the medical query. We here apply existing text retrievers, such as auto-merging or BM25, during this phase.

3.3 EHR-Based Report Ranker f_{Ranker}

EHR as a structured data organization, typically consists of multiple tabular data, each recording specific medical information about patients. In this study, we focus on measuring the similarity between patients using the following three medical entities:

- Diagnosis: Identified disease assigned to a patient, represented by ICD-10 code.
- Medication: Prescribed drugs administered to a patient, recorded using NDC code.
- Procedure: Medical intervention, operation, or clinical process performed on a patient, represented by ICD-10 code.

Quantify the similarity between patients based on these three dimensions offers a comprehensive criterion for identifying clinically relevant reports.

For a patient p, these three medical entities are represented as sets E_p^{Diag}, E_p^{Med}, and E_p^{Proc}, respectively. Given two patients p and p', we first compute the set similarity between them using each medical information:

$$\tau_{\text{Diag}} = f_{\text{similarity}}(E_p^{\text{Diag}}, E_{p'}^{\text{Diag}}), \tag{4}$$

$$\tau_{\text{Med}} = f_{\text{similarity}}(E_p^{\text{Med}}, E_{p'}^{\text{Med}}), \tag{5}$$

$$\tau_{\text{Proc}} = f_{\text{similarity}}(E_p^{\text{Proc}}, E_{p'}^{\text{Proc}}) \tag{6}$$

where $f_{\text{similarity}}(\cdot, \cdot)$ is a set similarity metric. We use the Jaccard Index in this study because each patient is represented as sparse sets of positive diagnosis, medication, and procedure codes, for which shared presences are more clinically informative than shared absences. This makes set-overlap similarity more appropriate and more interpretable than distances such as Euclidean or Manhattan distance over dense vectors. Finally, these similarity metrics are aggregated using a weighted sum:

$$\tau = \lambda_1 \tau_{\text{Diag}} + \lambda_2 \tau_{\text{Med}} + \lambda_3 \tau_{\text{Proc}} \tag{7}$$

where $\lambda_{1/2/3}$ is the hyperparameter balancing the importance of each metric. We perform pairwise similarity comparisons between the query patient and other patients within the EHR, returning the discharge summaries of the top-k most similar patients as results.

Efficiency Analysis. The overall computation is practically efficient since the computation of Jaccard Index can be significantly accelerated with some libraries, such as Faiss (Douze et al., 2024) and NumPy (Harris et al., 2020). Besides, indexing medical entities from tabular data enables fast lookups, further reducing computational overhead.

4 DISCHARGEQA Dataset

To evaluate LLMs' capability in utilizing the retrieved experience, we construct a medical question-answering dataset specifically designed for discharge-related queries based on MIMIC-IV (Johnson et al., 2023b). Each question in the dataset pertains to critical discharge information, including the patient's diagnosis, prescribed medications, and post-discharge instructions.

4.1 Dataset Introduction

Overview. As shown in Table 1, DISCHARGEQA consists of a total of 1,280 QA pairs, each associated with a patient ID and corresponding clinical background. The questions in DISCHARGEQA can be categorized into three main types:

	Task	Response Type	#Query	Example Question	Practical Significance	Background Source	Option Source
	Diagnosis Inference	Multi-select	436	"Which diagnoses should be documented in the patient's discharge summary?"	Reflects a doctor's process of identifying all relevant diagnoses based on clinical profile.	Clinical profile	Discharge Report & EHR
DISCHARGEQA	Medications Inference	Multi-select	444	"Which medications should be prescribed to the patient at discharge?"	Simulates the doctor's task of ensuring correct medications are prescribed based on hospital treatment.	Clinical profile & In-hospital progress	Discharge Report & EHR
	Instructions Inference	Single-select	400	"What is the best instruction for this patient?"	Mimics the final step of doctors' advising patients with appropriate post-discharge care instructions.	Clinical profile & In-hospital progress	Discharge Report & AI
EHRNoteQA	Clinical Inference	Single-select & Open-Ended	962	"What was the treatment provided for the patient's left breast cellulitis?"	Extract and answer based on content from full discharge notes	Full clinical notes	Discharge Report & AI
CliniQG4QA	Retrieval	Text Span	1,287	"Why has the patient been prescribed hctz?"	Retrieve the related content as answer from report	Full clinical notes	/

Table 1: Comparison of DISCHARGEQA and the previous EHR-related QA benchmarks.

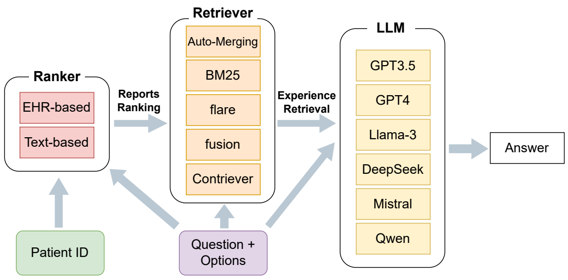

Figure 3: Inference pipeline of DISCHARGEQA.

- Diagnosis Inference: Questions related to identifying the patient's medical diagnosis.
- Medication Inference: Questions regarding the medications prescribed, including dosage, frequency, and purpose.
- Instruction Inference: Questions focused on discharge instructions, such as follow-up care, activity restrictions, and self-care guidelines.

These three categories collectively cover the key aspects of discharge-related patient care, requiring LLMs to perform non-trivial reasoning based on the given clinical background. Moreover, all these problems can potentially benefit from the retrieved experience, mimicking the way clinicians apply past clinical knowledge to make medical decisions during discharge.

Evaluation Settings. Each problem in DISCHARGEQA includes a problem description, the patient's clinical background, and multiple options for the LLM to choose from. While instruction inference uses a single-select setup, diagnosis and medication inference adopt a multi-select setup, requiring the model to select multiple options to answer the questions. Each option corresponds to a specific diagnosis or treatment. This multi-select format presents a more challenging and realistic setting for LLMs, as clinicians often need to identify and address multiple coexisting conditions.

The overall inference pipeline is presented in Figure 3, where we implement several components, including Ranker, Retrieval, and LLM agent, to support discharge-related QA using EXPRAG.

Comparison.

Compared with existing QA benchmarks that focus primarily on general clinical QA or information retrieval, DISCHARGEQA centers on the discharge procedure, simulating a doctor's medical reasoning process: inferring the diagnosis from the clinical profile, prescribing appropriate treatments, and summarizing the condition—offering a more realistic scenario. Additionally, we utilize EHR to generate contextually relevant options, requiring LLMs to perform non-trivial reasoning to solve the tasks. More details are provided in Table 1.

4.2 Dataset Construction

Patients Filtering. We first filter out low-quality patient records in MIMIC-IV for various reasons. Starting with all 430,000 patients, we remove encounters without available discharge summaries, leaving 320,000 patients. Next, we filter out patients with fewer than 3 or more than 40 entries in any of the diagnosis, medication, or procedure records, resulting in a final dataset of 28,000 patients for generating QA pairs. For instruction inference, we further exclude patients with excessively short discharge summaries using GPT-4o, as explained in A.1.

Background Generation. To enable LLMs to make realistic medical decision, it is necessary to

Model	Context	Instruction	Diagnosis		Medication	
		Acc(%)	Acc(%)	F1	Acc(%)	F1
Deepseek-R1-8B	Direct-Ask	70.0	10.78	0.363	0.69	0.217
	bge-small-en	72.0	12.84	**0.381**	1.58	0.230
	ExpRAG EHR	**75.3**	11.01	0.379	0.90	**0.241**
DeepSeek-V3.1	Direct-Ask	86.3	9.63	0.509	3.83	0.527
	bge-small-en	91.0	10.78	0.517	7.21	0.584
	ExpRAG EHR	**91.3**	11.24	**0.523**	8.33	**0.642**
Llama-4-17B	Direct-Ask	92.0	5.73	0.445	5.86	0.557
	bge-small-en	92.8	6.19	0.476	5.86	0.606
	ExpRAG EHR	**93.5**	7.57	**0.485**	8.68	**0.649**
Qwen3-30B-A3B	Direct-Ask	90.8	15.60	0.415	1.13	0.280
	bge-small-en	93.8	18.12	0.502	2.25	**0.366**
	ExpRAG EHR	<u>**95.3**</u>	17.43	**0.528**	1.35	0.355
BaichuanM2	Direct-Ask	86.0	11.24	0.520	2.70	0.447
	bge-small-en	86.8	15.37	0.407	3.60	0.513
	ExpRAG EHR	**87.8**	16.97	<u>**0.554**</u>	3.83	**0.549**
GPT-3.5	Direct-Ask	73.0	18.50	0.498	1.15	0.234
	bge-small-en	78.8	15.60	0.405	0.68	0.317
	ExpRAG EHR	**79.5**	18.81	**0.504**	1.80	**0.371**
GPT-4o	Direct-Ask	90.0	9.86	0.510	3.65	0.486
	bge-small-en	90.3	8.26	0.493	4.95	0.601
	ExpRAG EHR	**91.3**	<u>**21.33**</u>	**0.530**	9.68	**0.638**
GPT-5 nano	Direct-Ask	81.0	10.78	0.485	4.50	0.576
	bge-small-en	84.0	13.99	0.525	7.21	0.620
	ExpRAG EHR	**84.0**	16.06	**0.542**	10.78	**0.672**
GPT-5 mini	Direct-Ask	85.3	4.81	0.459	7.66	0.622
	bge-small-en	89.5	4.36	0.466	8.33	0.636
	ExpRAG EHR	**91.5**	5.05	**0.462**	<u>**11.71**</u>	<u>**0.678**</u>

Table 2: Performance Comparison Across Multiple LLMs using DischargeQA.

offer a clinical background of the patient along with the question. To **avoid label leakage** during the context generation, we propose leveraging the structured format of the discharge summary, which consists of the following components:

- Clinical profile: Essential patient demography, the presenting condition, and initial clinical assessments.
- In-hospital progress: The interventions, therapies, and the patient's clinical progress during hospitalization.
- Discharge plan summary: The details of discharge diagnosis and medication, and instructions for post-hospital care.

An illustrative example can be found in Appendix A.3. These three components represent the core of discharge decision-making procedure, aligning with the three main problem types in DischargeQA. Accordingly, we present the summarized sections preceding the questions as context. For example, the clinical profile serves as the background for diagnosis-related questions, while the in-hospital progress is additionally included for medication-related questions. Notably, basic patient demographic information is always included as part of the contextual background.

Option Generation. For diagnosis and medication inference, we utilize EHR to generate nontrivial candidate options by extracting all associated diagnoses and medications of a patient and feeding them into GPT-4o to select contextually relevant candidates, ensuring a challenging selection process. For instruction inference, GPT-4o first summarizes the key points of the ground-truth answers and then applies permutations to generate plausible yet incorrect candidate options. Notably, EHR tabular data usually has **very limited overlap** with our task for discharge, such as in-hospital medications, due to professionalism and use cases, have much limitation than discharge prescriptions. For all tasks, the correct options are directly extracted from the discharge reports.

5 Experiments

In this section, we evaluate ExpRAG on DischargeQA with nine LLMs, comparing it with the text-based report rankers. We also analyze the effects of balancing coefficients, the number of similar patients, and retriever choices, followed by case studies on the retrieved experience.

5.1 Comparison of LLMs

We evaluated the performance of 9 state-of-the-art LLMs of varying scales on three clinical tasks: discharge instructions, diagnosis, and medications. These models included close-source models GPT-3.5 (OpenAI, 2022), GPT-4o (OpenAI, 2024)), GPT-5-mini GPT-5-nano (OpenAI, 2025), four thinking models—Deepseek-R1-8B, Deepseek-V1 (Guo et al., 2025), Llama-4-17B(Maverick) (Meta AI, 2025) and Qwen3-30B-A3B (Team, 2025), and a medical model BaichuanM2 (Team et al., 2025).

Hyperparameters. The default balancing coefficients $\lambda_{1/2/3}$ are set to 1/3 each and auto-merging (Liu, 2022) is used as the retriever. The number of similar patients is set as 15. More details can be found at A.7.

Metrics. We report accuracy across all the tasks, calculated as the percentage of correctly answered questions. Note that for multi-select problems, a question is considered answered correctly only when all correct options are selected. We additionally report F1 scores for the multi-select problems to provide a more comprehensive analysis of LLM performance on these challenging tasks.

Results. Table 2 summarizes the performance of all LLMs with ExpRAG on the three clinical tasks. ExpRAG $_{\text{EHR}}$ delivers the strongest or tied-strongest results on most metrics and models. Additionally, the multi-select tasks (diagnosis and medication) prove significantly more challenging, as most LLMs achieve accuracy below 20%, indicating the limited medical reasoning capabilities of current large language models, the value of challenge of DischargeQA.

5.2 Comparison of Report Ranker

To verify the effectiveness of ExpRAG and the utilization of EHR, we implement baselines that perform report ranking solely based on text, i.e., a text-based ranker, as a key part in other traditional RAG methods. Using embedding model `bge-small-en-v1.5` (Xiao et al., 2023), a widely used compact sentence-embedding encoder due to its high correlation scores in Table 4, we embed queries (question, options, background which includes demographic and clinical-profile information) to be computed similarity with each discharge report embedding. The top-k similar reports are then retrieved, followed by a text retriever to extract the relevant information.

Rankers	Instruction	Diagnosis		Medication	
	Acc	Acc	F1	Acc	F1
bge-small-en	78.8	15.60	0.405	0.68	0.317
all-MiniLM-L6	79.3	17.43	0.476	1.13	0.316
paraphrase-L3	77.5	18.81	0.492	1.58	0.330
MedCPT	73.8	18.35	0.508	1.58	0.331
ExpRAG $_{\text{EHR}}$	**79.5**	**18.81**	**0.504**	**1.80**	**0.371**

Table 3: Performance Comparison with Embedding Models as Reranker vs EHR-based Reranker.

The results are presented in Table 2. We observe that ExpRAG outperforms the text-based ranker in most cases, achieving an average relative improvement of 5.2%. Notably, the EHR-based ranker leverages structured EHR data for ranking, eliminating the need for an embedding process and thereby enabling a more efficient pipeline. Because both ExpRAG and the text-based baselines retrieve from the same pool of discharge summaries, these gains reflect differences in how clinically similar patients are ranked rather than differences in the retrieval source itself. We therefore interpret these results as robust retrieval improvements on difficult tasks, rather than evidence of task saturation.

As shown in Table 3, we compare our EHR-based ranker with two additional sentence-embedding models: `all-MiniLM-L6-v2` and `paraphrase-MiniLM-L3-v2` (Reimers and Gurevych, 2019), and one strong domain encoder MedCPT (Jin et al., 2023) using GPT3.5 as backbone LLM (A.5.1). As a result, ExpRAG $_{\text{EHR}}$ outperforms all text-based variants in terms of all the metrics.

Retrieval Correlation Comparison. We conducted an experiment comparing patients similarity scores generated by the EHR-based method with those reliably annotated by LLMs. A higher correlation between the two indicates stronger retrieval performance by the EHR-based approach. Specifically, we randomly sample 100 target patients from DischargeQA, as detailed in Appendix A.8.

Rankers	Pearson	Spearman
bge-small-en-v1.5	0.639	0.623
all-MiniLM-L6	0.640	0.618
paraphrase-MiniLM-L3	0.478	0.481
ExpRAG EHR	**0.669**	**0.648**

Table 4: Retrieval Performance Comparison of Rerankers using Pearson and Spearman Correlation.

Model	Instruction	Diagnosis		Medication	
	Acc	Acc	F1	Acc	F1
Uniform	79.5	18.81	0.504	1.8	0.371
Task-focused	77.5	10.91	0.377	0.91	0.322
Complementary	76.8	18.18	0.446	2.73	0.305

Table 5: Performance Comparison for Coefficients Distributions DischargeQA.

Each target–candidate pair is scored by GPT4o-mini on the three modalities (diagnoses, procedures, prescriptions) with a single "overall" similarity. For each ranker (i.e., three text-embedding models and our EHR-based ExpRAG), we compute Pearson and Spearman correlation between its scores and annotations, and average over all 100 targets (Table 4). Over the strongest text-embedding baselines,the results by ExpRAG confirm that explicit EHR similarity provides more faithful retrieval signals than generic sentence embeddings.

5.3 Further Analysis

We conducted additional studies to investigate the impact of key components on the performance of ExpRAG, including the number of similar patients k, and the balancing coefficients $\lambda_{1/2/3}$. GPT-3.5 is applied as the backbone by default.

Balancing Coefficients. As shown in Eq. 4, we introduce three coefficients to balance the similarity computed based on diagnosis, medication, and procedures. We apply an equal distribution by default and explore the effect of using different weighting strategies here:

- Task-focused weighting: Assign a weight of 1 to the task-relevant similarity measure and 0 to the others. For example, $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 0$ for diagnosis inference.
- Complementary weighting: Assign a weight of 1 to the two less relevant similarity measures while setting the task-relevant measure to 0. For example, $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 1$ for diagnosis inference.

# Similar Patients	Instruction	Diagnosis		Medication	
	Acc	Acc	F1	Acc	F1
$k = 5$	80.00	19.04	0.511	1.80	0.366
$k = 10$	78.75	18.58	0.511	1.35	0.371
$k = 15$	79.50	18.81	0.504	1.80	0.371
$k = 20$	80.75	19.50	0.524	1.80	0.377
$k = 25$	82.25	19.27	0.515	1.35	0.352

Table 6: GPT-3.5 performance with different k.

The results are shown in Table 5. We can find that uniform weighting can achieve the best performance in most cases, demonstrating that information from multiple clinical dimensions can provide a more comprehensive context.

Top-k Patients. We vary the number of retrieved similar patients k on QA performance. As shown in Table 6, different tasks have varies trends. For example, on Instruction task, accuracy starts to increases with larger number. While for Diagnosis and Medication tasks, beyond $k = 20$, performance fluctuates, suggesting that while more retrieved candidates provide useful context, excessive retrieval may introduce irrelevant or conflicting information, leading to slight declines in accuracy.

5.4 Case Study

We perform a focused case study on a Discharge Diagnosis task from DischargeQA. Specifically, we analyze a patient (ID: 20453584) presenting with bilateral ulnar paresthesias and neck pain. Similar patients are identified by matching ICD/NDC codes from structured EHR data (Figure 8). Reviewing discharge summaries of these similar patients revealed shared key diagnostic features—including cervical disc herniation, spinal stenosis, and upper extremity neurological symptoms—which substantially clarified the target patient's clinical picture. For instance, similar patients exhibiting C6-C7 disc herniations and spinal stenosis provided critical evidence, improving the interpretation of the target patient's symptoms and supporting a more accurate final diagnosis. We elaborate the explanation in Appendix A.9.

6 Conclusion

Inspired by the importance of experience in clinical decision-making, we propose a novel coarse-to-fine retrieval framework, EXPRAG, to utilize knowledge from similar patient records. Specifically, we introduce EHR as a knowledge basis and employ a reliable similarity measurement algorithm to narrow down the candidate pool to have relevant and useful content. Evaluated on DISCHARGEQA, EXPRAG consistently improves the performance of various LLMs, highlighting the potential of leveraging past experience to enhance model performance.

Limitations

While EHR provides abundant medical information, such as lab test results, our proposed EXPRAG currently utilizes only diagnosis, medication, and procedures as an initial exploration, which provides valuable insights and promising directions for future research. In addition, our evaluation is conducted within a single MIMIC-IV-derived cohort, where both query patients and retrieval candidates come from the same institutional corpus. This matches the intended deployment setting of within-institution EHR retrieval, but it does not establish cross-dataset or cross-institution generalizability. Extending EXPRAG to additional EHR datasets and multi-center cohorts is an important direction for future work. Finally, DISCHARGEQA currently consists solely of multi-option questions, which can be enhanced to be open-ended to comprehensively evaluate the generative capabilities of LLMs, When more economical and accurate LLMs are developed.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2403317.

References

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097, Marseille, France. European Language Resources Association.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Jacob Devlin, Ming-Wei Chang, et al. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv:2401.08281*.

Jungwei Fan. 2019. Annotating and characterizing clinical sentences with explicit why-qa cues. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 101–106.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *AAAI*, volume 32.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Tinglin Huang, Syed Asad Rizvi, Rohan Krishna Thakur, Vimig Socrates, Meili Gupta, David van Dijk, R Andrew Taylor, and Rex Ying. 2024. Heart: Learning better representation of ehr data with a heterogeneous relation-aware transformer. *Journal of Biomedical Informatics*, 159:104741.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11).

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Alistair E. W. Johnson et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. In *Proceedings of NeurIPS 2024 (Datasets and Benchmarks)*.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. Ehrsql: A practical text-to-sql benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35, pages 15589–15601. Curran Associates, Inc.

Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. 2025. RGAR: Recurrence generation-augmented retrieval for factual-aware medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4006–4033, Suzhou, China. Association for Computational Linguistics.

Jerry Liu. 2022. LlamaIndex.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, et al. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2025. Llama 4 — maverick (17b, 128e) model card and release notes. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/` and model card on Hugging Face. Model id: Llama-4-Maverick-17B-128E (instruction / instruct variants exist).

Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, Jungwei Wilfred Fan, et al. 2023. Extractive clinical question-answering with multianswer and multifocus questions: Data set development and evaluation study. *JMIR AI*, 2(1):e41818.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025. Gpt-5 system card (includes gpt-5-mini and gpt-5-nano). `https://openai.com/index/gpt-5/` and System Card PDF. Model names: gpt-5, gpt-5-mini, gpt-5-nano.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yifei Shen, Yilun Zhao, Justice Ou, Tinglin Huang, and Arman Cohan. 2026. Patient-similarity cohort reasoning in clinical text-to-SQL. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1367–1412, Rabat, Morocco. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. RadQA: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France. European Language Resources Association.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, and et al. Yijie Zhou. 2025. Baichuan-m2: Scaling medical capability with large verifier system.

Qwen Team. 2025. Qwen3.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. 2025a. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. *arXiv preprint arXiv:2510.02328*.

Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. 2025b. A survey of large language models for text-guided molecular discovery: from molecule generation to optimization. *arXiv preprint arXiv:2505.16094*.

Shitao Xiao, Zheng Liu, Peitian Zhang, et al. 2023. C-pack: Packaged resources to advance general chinese embedding. *arxiv:2309.07597*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions.

Lee Xiong, Chenyan Xiong, Ye Li, et al. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2.

Haohan Yuan, Sukhwa Hong, and Haopeng Zhang. 2026. Strucsum: Graph-structured reasoning for long document extractive summarization with llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3708–3721.

Haohan Yuan and Haopeng Zhang. 2025. Understanding llm reasoning for abstractive summarization. *arXiv preprint arXiv:2512.03503*.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 580–587. IEEE.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1):909.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2024. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–26.

A Appendix

A.1 DISCHARGEQA

To evaluate LLMs in real-world clinical scenarios, we constructed a dataset combining structured tables from MIMIC-IV and unstructured discharge notes from MIMIC-IV-note, totaling over 140 000 patient records. The whole process is shown in 4

Patient selection. As in 4.2, a valid pool of candidates for target or similar patients is the key point for retrieving information from similar patients.

Starting from the 430 k patients in MIMIC-IV, we retain (i) encounters with a discharge summary, (ii) 3–40 rows in each of `diagnoses`, `prescriptions`, and `procedures`, and (iii) discharge notes that contain ≥ 4 instruction bullet points identified by GPT-4o. This yields 28 000 admissions, each can give enough structured items and narrative content to form challenging QA pairs, so that there are enough information to compare to other patients, and to have candidate options as wrong answer if there is no overlap in discharge reports.

Note segmentation and exposure. Each discharge summary is heuristically split into seven sections *(Patient Demography, Presenting Condition, Clinical Assessment, Treatment Plan, In-Hospital Progress, Discharge Summary, Post-Discharge Instructions)* and mapped to three temporal phases: **pre-diagnosis**, **in-hospital**, and **post-discharge**, matching the 3 sections **Clinical profile**, **In-hospital**, and **Discharge Plan** as Figure 6. For every task we reveal only the phases that would have been available to the clinician at decision time, preventing label leakage.

Problem Design with EHR

- **Golden answers** are clinician-authored discharge–note items.
- **Distractors** are drawn from the same patient's structured tables (drugs for medication, ICD-coded diagnoses for diagnosis).
- Overlaps between structured candidates and golden answers are merged via GPT-4o.
- Unlike single-choice formats, both diagnosis and medication tasks use *multi-select*, requiring selection of *all* correct items, thus raising task difficulty.

Construction Prompt Design Prompts to generate options are carefully crafted to combine EHR tabular data and golden answer from discharge note. As an example, the prompt as Figure 7 defines the role of the model as a clinician and provides three key components: the list of discharge diagnoses, a database of historical diagnoses, and summarized background information extracted from the discharge summary. We keep this summarization step tightly constrained to faithful compression of clinically available information, aligning recent work which suggests that stronger reasoning in summarization does not necessarily improve factual faithfulness(Yuan and Zhang, 2025). The task requires the model to identify which diagnoses should be included in the discharge summary by reasoning through the given data. Correct options are derived from the discharge diagnosis list, while incorrect but plausible options are generated from the diagnoses database. To ensure realism, GPT-4o is used to handle overlap, summarize long diagnoses, and align outputs with clinical expectations, providing a rigorous framework for evaluation.

Dataset Demographics

Top 10 Diagnoses. Table 7 for diagnoses.

Top 10 Prescriptions. Table 8 for medications.

Top 10 Procedures. Table 9 for procedures.

Gender distribution. See Table 10.

Age distribution (5-year bins). See Table 11.

A.2 Compare Two Rankers using similar patients

Shown as Figure 5.

A.2.1 Conventional RAG

A traditional RAG pipeline for clinical question answering consists of:

- Encoding a query (e.g., a discharge planning question) into an embedding vector.
- Chunking a large corpus into smaller text passages and embedding each chunk.
- Finding top-k relevant text chunks by comparing similarity scores between query embeddings and chunk embeddings.
- Retrieving the most relevant text passages.
- Feeding the retrieved text to the LLM for answer generation.

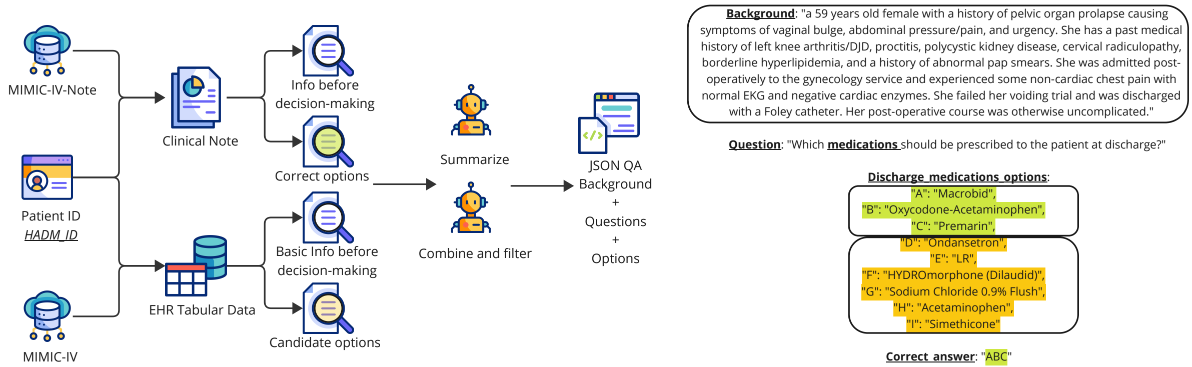

Figure 4: DischargeQA generation workflow.

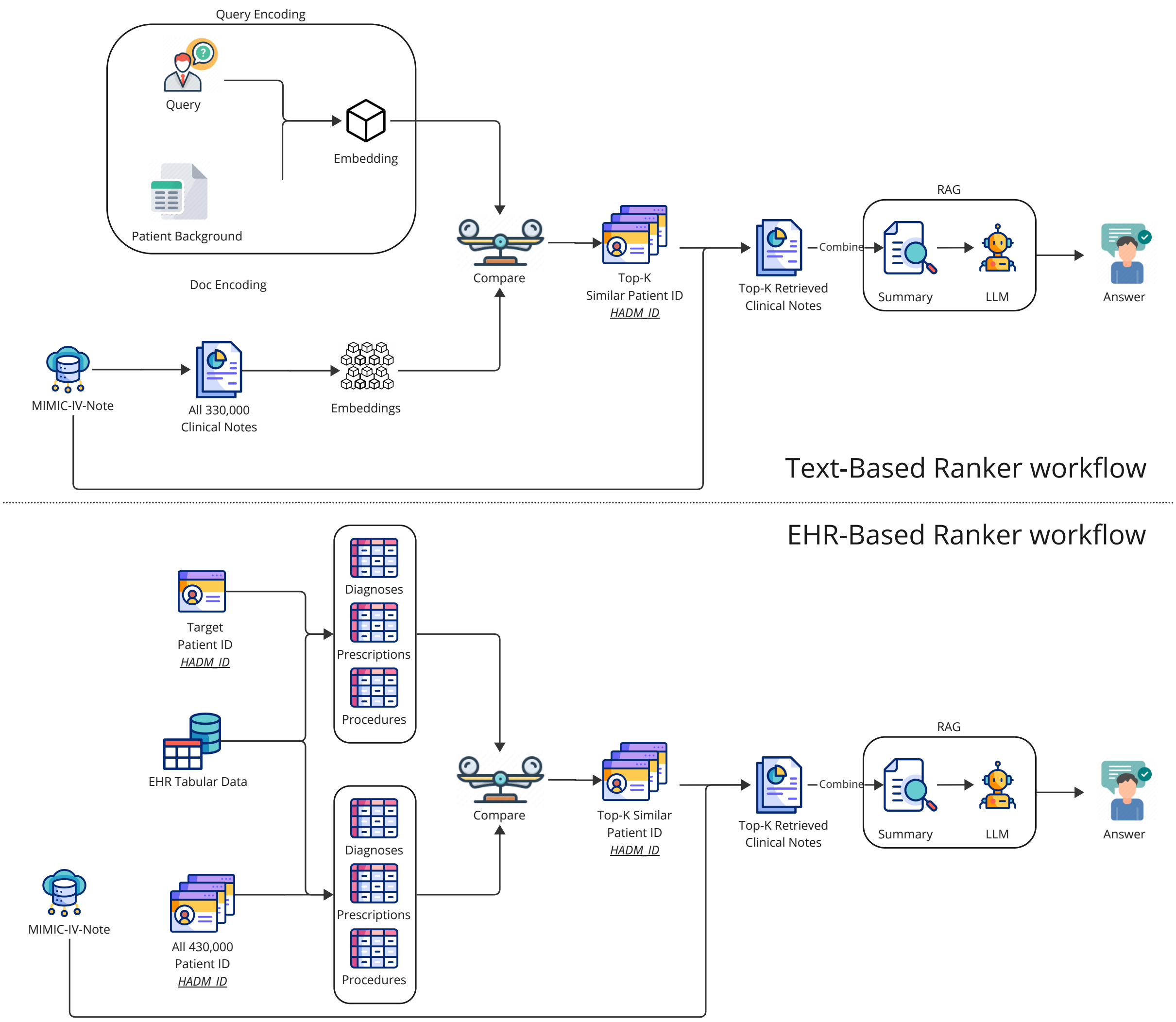

Figure 5: Text-Based Ranker vs EHR-Based Ranker workflow.

ICD Code	Diagnosis	Count
4019	Unspecified essential hypertension	238
2724	Other / unspecified hyperlipidemia	180
41401	Coronary atherosclerosis of native coronary artery	136
53081	Esophageal reflux	108
25000	Type II diabetes, no complication	92
E785	Hyperlipidemia, unspecified	67
3051	Tobacco-use disorder	64
I10	Essential (primary) hypertension	63
4280	Congestive heart failure, unspecified	56
2449	Unspecified acquired hypothyroidism	53

Table 7: Top-10 discharge diagnoses in DISCHARGEQA.

Drug Code	Medication	Count
NACLFLUSH	Sodium Chloride 0.9% Flush	665
DOCU100	Docusate Sodium	470
HEPA5I	Heparin 5 000 U/mL (IV)	430
ACET325	Acetaminophen 325 mg	383
SENN187	Senna	340
ONDAN4I	Ondansetron 4 mg IV	324
APAP500	Acetaminophen 500 mg	212
OXYC5	Oxycodone 5 mg IR	191
BISA10R	Bisacodyl 10 mg PR	150
LRLV	Lactated Ringer's	145

Table 8: Top-10 discharge prescriptions.

However, applying this approach to EHR discharge summaries presents challenges: **Inefficiency**: The MIMIC-IV discharge summary corpus is over 4GB in size, making fine-grained chunk retrieval computationally expensive. **Loss of Context**: Traditional chunking disrupts the continuity of patient history, making it difficult for LLMs to infer longitudinal medical decisions. **Scattered Information:** Important information may be spread across multiple notes, making individual paragraph retrieval suboptimal. This limitation is consistent with recent findings in long-document summarization showing that LLMs often struggle to preserve document structure and identify salient information in lengthy inputs, motivating structure-aware methods for handling long clinical reports (Yuan et al., 2026)

A.2.2 Text-Based Ranker

Instead of retrieving individual paragraphs or sentences, we treat each patient's discharge summary as a retrievable document and adapt the RAG pipeline accordingly:

Query Encoding Convert the question, options and target patient's background into an embedding vector.

Patient-Level Indexing Store each patient's discharge summary as a separate retrievable document and compute its embedding.

Find Similar Patients Using Embeddings Rank top-k similar patients from all patients based on similarity between the query embedding and patient embeddings.

Retrieve Top-K Patient Summaries Select the top-k most similar patient discharge summaries as a document source, like "experience".

Run RAG on Retrieved Summaries Use the retrieved summaries as references for downstream LLM-based answer generation, aligning with the broader trend of domain-specialized LLM systems that use task-guided external information in scientific applications (Wang et al., 2025b). Traditional RAG is running by embedding the query and the document source, retrieving chunks as reference, and answering query to generate the correct answer.

By preserving full discharge summaries, Text-based retrieval maintains patient-level coherence while improving retrieval efficiency.

ICD-9 Code	Procedure	Count
0040	Procedure on single vessel	111
8856	Coronary arteriography (two catheters)	91
0066	Percutaneous coronary angioplasty (PTCA)	84
3722	Left-heart cardiac catheterisation	74
8938	Non-operative respiratory measurements	73
0045	Insertion of one vascular stent	71
3899	Other puncture of vein	71
3607	Drug-eluting coronary-artery stent(s)	65
8952	Electrocardiogram	59
8744	Routine chest X-ray	52

Table 9: Top-10 in-hospital procedures preceding discharge.

Gender	Count
Female (*F*)	330
Male (*M*)	397

Table 10: Gender distribution in DISCHARGEQA.

Age Range	Count
15–19	2
20–24	14
25–29	21
30–34	38
35–39	34
40–44	48
45–49	55
50–54	80
55–59	86
60–64	74
65–69	74
70–74	63
75–79	65
80–84	39
85–89	19
90–94	15

Table 11: Age distribution in DISCHARGEQA.

A.2.3 EHR-Based Ranker

To further improve retrieval relevance, we introduce an EHR-based Ranker that utilizes structured patient data for retrieving better discharge summaries as the "experience" document source:

Identify the Target Patient's Condition Extract structured EHR tabular data (ICD codes for diagnoses, prescriptions, and procedures) from the target patient.

Find Similar Patients Using Structured EHR Data Compute similarity between the target patient's structured data and all other patients based on shared ICD codes in MIMIC-IV, after scanning all recorded patient data. Jaccard index is used to calculate the similarity.

Retrieve Top-K Similar Patients Select the top-k most similar patients based on their structured medical records.

Extract Their Discharge Summaries Retrieve the corresponding discharge summaries for the top-k similar patients.

Run RAG on Retrieved Summaries Feed the retrieved summaries into an LLM for answer generation, same as the last step in Text-Based Ranker.

By incorporating structured EHR data, this method ensures that retrieval is clinically relevant, going beyond semantic text similarity, echoing recent medical knowledge augmentation work (Wang et al., 2025a). This process resembles an experience-based search: identifying similar past experiences and adapting them to solve the problem at hand.

A.3 Dataset Details

MIMIC-IV[2] This database (Johnson et al., 2016, 2023b,a) contains information on over 40,000 patients admitted to the critical care units at Beth Israel Deaconess Medical Center from 2001 to 2012 and has been widely used in prior research. The database is publicly available for research purposes, with strict de-identification protocols to protect patient privacy, making it a valuable resource for developing and evaluating machine learning models in healthcare. The data is hierarchically organized, with each patient record comprising multiple encounters, each containing various entities such as demographics, medications, diagnoses, procedures, and lab results. Additionally, the database includes

[2]https://mimic.mit.edu/docs/iv/

Retriever	Instruction Acc
Auto Merging	79.5
Sentence Window	74.5
BM25	68.0
BM25+	69.0
Flare	74.5
Contriever	69.0

Table 12: Performance Comparison of Retrievers.

unstructured data, such as discharge reports and X-ray images, with each admission marked by a date and timestamp.

Structure of Discharge Report As shown in Figure 6, a discharge summary follows a structured format that provides a comprehensive overview of a patient's hospitalization and care journey. It is typically divided into three main parts: (1) The Clinical Profile, which includes essential patient information, presenting conditions, and the initial clinical assessment upon admission; (2) The In-Hospital Progress, which documents the treatment plan, administered therapies, and the patient's progress throughout the hospital stay; and (3) The Discharge Plan Summary, which summarizes the patient's discharge status, prescribed medications, and detailed post-discharge instructions for ongoing care.

A.4 Retriever and Experiments

After the initial ranking by EXPRAG, contents from discharge summaries are retrieved by retrievals. As part of the ablation study, using GPT-3.5 as LLM backbone and EXPRAG as ranker, our experiments for text retrievers include: Auto-merging (Liu, 2022), sentence-window (Liu, 2022), BM25 (Robertson et al., 2009), BM25+ (the combination of BM25 and word embeddings), Contriever (Izacard et al., 2021), and flare (Jiang et al., 2023). As shown in Table 12, auto-merging achieves the best performance, and it's used as the default text retriever. However, Contriever, which utilizes unsupervised dense representations, also underperforms in our cases, suggesting the need for medical domain-specific fine-tuning.

Auto-merging Auto-merging retrieval in RAG by LlamaIndex (Liu, 2022) hierarchically structures documents into parent and child nodes, allowing for the retrieval of larger, more coherent context by merging child nodes into parent nodes when multiple related chunks are relevant to a query.

Sentence-window Sentence-window retrieval by LlamaIndex (Liu, 2022) parses documents into individual sentences with surrounding context, enabling fine-grained retrieval while maintaining local coherence by including a window of adjacent sentences.

BM25 & BM25+ BM25 (Best Match 25) is a ranking function used in information retrieval that scores documents based on the frequency of query terms within them, taking into account document length and term frequency saturation. BM25+ by LlamaIndex (Liu, 2022) combines retrieval methods BM25 and vector-based retrieval, to leverage the strengths of both approaches. This hybrid technique allows for capturing both keyword relevance and semantic similarity, often using algorithms like Relative Score Fusion to re-rank and merge results from different retrievers, resulting in more accurate and comprehensive search outcomes.

Flare FLARE (Forward-Looking Active REtrieval) enhances RAG by enabling the language model to anticipate future content needs, iteratively predicting upcoming sentences and retrieving relevant information when encountering low-confidence tokens, thus improving response accuracy and contextual relevance.

Contriever Contriever is a single-tower dense retrieval model that employs self-supervised contrastive learning to enhance document embeddings for retrieval tasks. It encodes both queries and documents using the same encoder, producing dense vector representations. The model utilizes a self-supervised contrastive learning approach with a loss function that optimizes embeddings by comparing relevant passages to negative (irrelevant) ones.

A.5 LLM Backbone

A.5.1 Proprietary Models

We selected GPT-3.5, GPT-4o, GPT-5-mini, GPT-5-nano as representative models due to their extensive real-world adoption and practical relevance demonstrate the applicability of our framework in everyday clinical settings. Exact model-id strings, decoding strategies are recorded in Table 13 for reproducibility.

Reason of GPT-3.5 for all ablations We run a controlled ablation with the LLM backbone fixed to GPT-3.5-turbo-1106. Holding the backbone constant ensures that all experiments use identi-

------------------------------Section 1 : Clinical profile------------------------------

[-----------------------Patient Demography-----------------------]
Name: ___ Admission Date: ___ Discharge Date: ___
Date of Birth: ___ Sex: F
Service: Surgery
Allergies: Penicillins, ACE Inhibitors
Attending: ___

[-----------------------Presenting Condition-----------------------]
Chief Complaint: Patient admitted with nausea and abdominal pain.
History of Present Illness: Patient presented with a history of nausea, vomiting, and abdominal pain for several days. Laboratory evaluation was unremarkable. CT scan showed a small bowel obstruction secondary to a Richter's hernia.
Past Medical History: HTN, HLD, Asthma, Tenosynovitis, Obesity.
Social History: ___
Family History: Non-contributory.

[-----------------------Clinical Assessment-----------------------]
Physical Exam:
- Vital signs: Temp 97.8°F, BP 130/65, HR 72, RR 16.
- Abdomen: Distended, possible tender mass below the umbilicus without erythema or guarding.

Pertinent Results:
- Labs: WBC 8.8, Hgb 13.1, Hct 37.7.
- CT Abdomen: Richter's hernia causing small-bowel obstruction, no perforation.

------------------------------Section 2: In-Hospital Progress------------------------------

[-----------------------Treatment Plan-----------------------]
Major Surgical or Invasive Procedure: Exploratory laparotomy, lysis of adhesions, hernia repair with mesh.

[-----------------------In-Hospital Progress-----------------------]
Brief Hospital Course:
Patient underwent surgery for small bowel obstruction secondary to Richter's hernia. Postoperative recovery was uneventful. Diet was gradually advanced, and pain was well-controlled.
Medications on Admission: Fluticasone, HCTZ, Naproxen, Prochlorperazine, Valsartan.

------------------------------Section 3: Discharge Plan------------------------------

[-----------------------Discharge Summary-----------------------]
Discharge Medications:
1. Oxycodone-Acetaminophen for pain, as needed.
2. Docusate Sodium, twice daily.
3. Ciprofloxacin, twice daily for 2 weeks.

Discharge Disposition: Home with services.
Discharge Diagnosis: Small bowel obstruction due to Richter's hernia.
Discharge Condition: Stable.

[-----------------------Post-Discharge Instructions-----------------------]
Discharge Instructions:
- Avoid driving or operating machinery while taking pain medications.
- Resume regular home medications.

Call your doctor or visit the ER if:
- You have chest pain, persistent vomiting, dehydration, or a fever above 101.5°F.
- There is redness, swelling, or drainage from the incision.

Activity:
- No heavy lifting for 6 weeks. Moderate exercise allowed, no abdominal exercises.

Wound Care:
- Shower allowed, no baths or swimming. Steri-strips will fall off on their own.

Follow-Up Instructions: ___

Figure 6: An example of discharge report, which can be split into 3 sections and 7 subsections: clinical profile, in-hospital progress, and discharge plan summary. Note: Some pertinent results from exams are before diagnosis, while some are after diagnosis or after procedures.

cal prompts and decoding settings, which reduces variance attributable to model-specific generation dynamics and focuses the comparison on the retrieval / ranking stage. Practically, using GPT-3.5 keeps compute and cost requirements modest while providing stable, predictable output formatting that benefits our automatic evaluation pipeline. Importantly, the improvements produced by the ranker are consistent with the trends observed when using alternative backbones (e.g., GPT-4o), indicating that the ranker's benefit generalizes beyond the fixed-backbone setting. Together, this ablation provides a clean, reproducible estimate of the ExpRAG ranker contribution under tightly controlled generation conditions.

GPT-3.5 & GPT-4o GPT-3.5, developed by OpenAI and released in November 2022, was followed by GPT-4o, also created by OpenAI and launched on May 13, 2024, marking a significant advancement with its ability to process and generate outputs across text, audio, and image modalities in real time.

GPT-5-mini & GPT-5-nano OpenAI released the GPT-5 family for developers on August 7, 2025, offering three API sizes—'gpt-5', 'gpt-5-mini', and 'gpt-5-nano'—to let developers trade off performance, cost, and latency. In this work we used the 'gpt-5-mini' and 'gpt-5-nano' endpoints via the OpenAI API.

LLM	Temperature	Top-p
GPT-3.5-turbo-1106	0	1.0 (Default)
GPT-4o-2024-11-20	0	1.0 (Default)
GPT-4o-2024-11-20	0	1.0 (Default)
GPT-5-mini-2025-08-07	0	1.0 (Default)
GPT-5-nano-2025-08-07	0	1.0 (Default)
DeepSeek-R1-8B	0	1.0
DeepSeek-V3.1	0	1.0
Llama-4-17B Maverick	0	1.0
Qwen3-30B	0.6	0.95

Table 13: Exact model ID and decoding hyperparameters. For closed-source models (e.g., GPT-3.5, GPT-4o), we use deterministic decoding (temperature = 0, default top-p) to control variance. For open models, we report the full decoding profile (temperature, top-p).

A.5.2 Open-source Models

Qwen3, Deepseek-R1, Deepseek-V3.1 and Llama4 were included as leading open-source models to benchmark the generalizability and robustness of our approach. They are all thinking models, which is an essential feature to tackle challenges in our tasks.

Deepseek-R1-8B DeepSeek-R1 is a powerful 671 billion parameter language model developed by DeepSeek AI, from which we use DeepSeek-R1-Distill-Llama-8B was derived as a more efficient 8 billion parameter version, distilled from the original model's knowledge to offer improved performance on reasoning tasks while maintaining computational efficiency

Deepseek-V3.1 DeepSeek-V3.1 is a hybrid, instruction-tuned large language model from DeepSeek AI that supports two inference modes: "Think" (chain-of-thought / stepwise reasoning) and "Non-Think" (fast answer mode), selectable via prompt templates or the vendor "DeepThink" toggle. The release emphasises improved agent skills and tool calling, with post-training optimizations to speed thinking and strengthen multi-step tool usage; we used the official model card and API documentation to select the 'Think' variant where multi-step reasoning was required.

Qwen3-30B-A3B Qwen3-30B-A3B is developed by Alibaba Cloud, released at April 2025. It is a dense-and-MoE model that has a "thinking mode" for deep reasoning, math, and coding. It outperforms earlier Qwen generations on logical reasoning, code generation, tool-using agent tasks, and human-preference benchmarks, while supporting 100 + languages for instruction following and translation. In our work we leverage its thinking mode to maximize accuracy on complex reasoning tasks.

Llama-4-17B Maverick Llama-4 Maverick (17B, MoE) is Meta's instruction-tuned, natively multimodal Mixture-of-Experts model designed for high-complexity reasoning and image+text tasks; the Maverick family (early fusion + MoE) targets improved reasoning, larger effective capacity, and efficient multimodal early-fusion workflows. In our experiments we used the Maverick 17B instruct variant and followed Meta's model-card prompt templates to elicit step-wise reasoning for tasks requiring intermediate chain-of-thought output.

A.5.3 Medical Models

Baichuan-M2 Baichuan-M2 is a 32-billion-parameter, medical-enhanced reasoning model developed by Baichuan AI that is specifically optimized for real-world clinical reasoning tasks. The model is trained within a novel Large Verifier System (including a Patient Simulator and Clinical Rubrics Generator) and uses a multi-stage reinforcement learning strategy (improved Group Relative Policy Optimization) to align interactive, dynamic medical decision-making with clinical evaluation metrics. On the HealthBench suite, Baichuan-M2 achieves state-of-the-art results among open-source models and narrows the gap with leading closed-source systems; model checkpoints and documentation have been publicly released to encourage reproducibility and follow-on research (Team et al., 2025).

Baichuan-M1 Trained from scratch on 20 T tokens that mix high-quality clinical and general texts, Baichuan-14B-M1 is the first open-source 14 B-parameter LLM purpose-built for medicine. It incorporates specialised heads, enabling fine-grained reasoning that matches general-domain peers on standard benchmarks yet surpasses models 5× larger on medical tasks. An updated architecture with longer-context handling further improves comprehension of lengthy clinical narratives and complex patient histories.

We have experiments shown in 14. Baichuan-14B-M1 exhibits the same limitations we observed in other compact models. When additional context from EHR- or text-based retrieval is supplied, the model occasionally confuses the target patient with the retrieved similar cases. Under RAG settings, this confusion produces > 5 % invalid responses—outputs that either fail our answer-parsing patterns or omit a decisive answer—leading to a noticeable overall drop in accuracy. Despite these invalid cases, Baichuan-M1 using EHR-Based EXPRAG still surpasses other approaches on the Medication task, underscoring its strength in pharmacological reasoning even when other aspects falter.

UltraMedical Llama-3-8B-UltraMedical is derived from Meta's Llama-3-8B and further tuned on the 410 K-example UltraMedical dataset (synthetic + curated), an 8 B-parameter biomedical specialist. It tops MedQA, MedMCQA, PubMedQA and MMLU-Medical, handily beating larger general models such as GPT-3.5 and Meditron-70B, and rivals domain-tuned Flan-PaLM and OpenBioLM-8B. The model targets exam-style question answering, literature comprehension and clinical-knowledge retrieval, making advanced medical NLP accessible at modest compute cost.

UltraMedical's 70 B variant has a 8 k-token context window. When we integrated the model into our retrieval-augmented generation pipeline, more than 50 % of the assembled prompts—target patient record + similar-patient excerpts + task instruction—were longer than 8 k tokens. These over-length inputs had to be truncated or rejected, which systematically stripped clinical details from half of our queries and degraded answer quality. Because this hard context limit blocked reliable end-to-end evaluation, we discontinued UltraMedical-70B for our study despite its otherwise strong domain results.

A.6 Prompt Templates

Figure 7 presents the prompt used to generate Diagnosis task options, which is the same way as generating options for Medication task.

A.7 Detailed Setup for Hyper-parameter selection

We fixed a single global configuration *a priori* for all tasks and models: (i) λ uses a symmetric prior (equal weights over diagnoses, medications, and procedures) to avoid bias and per-task / per-model cherry-picking; and (ii) k is chosen via a context-budget vs. marginal-gain trade-off. To verify that our results do not hinge on hidden tuning, we report sensitivity analyses (varying λ distributions and sweeping $k \in \{5, 10, 15, 20, 25\}$); scores vary modestly and method rankings remain unchanged, so we present a single conservative setting in the main tables to preserve comparability and reproducibility.

A.8 Detailed Setup for Ranker Evaluation

Patient and Candidate Sampling. We first randomly select 100 target patients from the MIMIC-IV cohort. For each target, we construct 100 candidate patients by:

- Sampling 20 uniformly at random from the filtered step described in 4.2 from full dataset (to include unrelated cases).
- Sampling 80 from a restricted pool formed by taking the patients (per target) with non-zero EHR similarity in either of three modalities (diagnoses, procedures, prescriptions).

Annotation with GPT-4o-mini. Each target–candidate pair is judged by GPT4o-mini along the three modalities. We average these three scores to obtain a single "ground truth" similarity for correlation.

Ranking Methods.

- **Text embeddings:** Cosine similarity over discharge summaries using three pretrained models (`bge-small-en-v1.5`, `all-MiniLM-L6-v2`, `paraphrase-MiniLM-L3-v2`).
- **EHR-based EXPRAG:** Jaccard similarity on code sets (`ICD/NDC`) for each modality, aggregated with equal weights.

Correlation Metrics. For each ranker and target patient, we produce a ranked list of 100 candidates. We then compute Pearson's ρ and Spearman's τ between the ranker's scores and the GPT4o-mini annotations, and average each metric across all 100 targets.

Results. Table 4 (main text) reports the final average correlations, demonstrating that EHR-based EXPRAG best aligns with human-like judgments.

A.9 Details in Case Studies

We conduct a case study focusing on Discharge Diagnosis in DISCHARGEQA. We examine the discharge diagnosis of a target patient (ID: 20453584) who presented with bilateral ulnar paresthesias and neck pain. We compare this patient with similar

Model	Context	Instruction	Diagnosis	Medication
		Acc(%)	**F1**	**F1**
Baichuan-M1	Direct-Ask	89.8	0.381	0.256
	Text-based	87.8	0.377	0.277
	EXPRAG $_{\text{EHR}}$	86.3	0.373	**0.285**

Table 14: Performance from Baichuan-M1, with 5% invalid answer on Text-based and EHR-based.

patients, who are selected by comparing ICD/NDC codes from EHR tabular data. Fig8 shows one example question in DISCHARGEQA and the similarities in the discharge reports between the target patient and the similar patients with IDs 25633130, 29378221 and 28817667. Upon reviewing the discharge summaries of the similar patients, it became clear that several key diagnostic features were shared with patient 20453584:

- **Disc Herniation**: Both the target patient and similar patients had disc herniations, with the target patient experiencing a C5-6 disc-osteophyte complex and the similar patients exhibiting C3-C4 and C6-C7 herniations.
- **Spinal Stenosis**: Many of the similar patients displayed **spinal stenosis**, which was consistent with the target patient's symptoms of narrowing of the spinal canal and foraminal narrowing.
- **Upper Extremity Symptoms**: The target patient reported **bilateral ulnar paresthesias**, which mirrored the bilateral symptoms observed in several similar patients, such as neck pain radiating to the arms and tingling in the extremities.

Results: By comparing the discharge summaries, key features from similar patients that influenced the diagnosis of the target patient:

- Similar patients with **C6-C7 disc herniation** and **radiculopathy** helped to refine the target patient's diagnosis, suggesting that similar nerve root involvement could explain the upper extremity symptoms.
- The presence of **spinal stenosis** and **neural foraminal narrowing** in several patients guided the understanding of the target patient's potential nerve compression, which contributed to the diagnosis of **spinal stenosis**.

The comparison to similar patients led to a more precise discharge diagnosis for the target patient, which included a C5-6 disc-osteophyte complex with associated **spinal canal narrowing** and **neural foraminal narrowing**. These insights allowed the LLMs to confirm the target patient's diagnosis, which aligned with options **A** and **G** — **disc osteophyte** and **spinal stenosis**.

A.10 Frequently-Asked Questions

A.10.1 Q1 — What is novel about this work?

Beyond the framework itself, we make the following contributions.

The DISCHARGEQA dataset.

- Existing public medical QA sets (e.g., MedMCQA, PubMedQA) are *general-purpose*; they do not target the discharge workflow.
- Typical EHR benchmarks emphasise single-modality predictions such as readmission or mortality. Crucial real-world decisions such as prescribing, coding diagnoses, and issuing discharge instructions remain under-explored.
- The few discharge-related corpora that do exist frame every query as single-choice information retrieval; they measure reading-time reduction, not clinical *reasoning* (see Table 1).
- DISCHARGEQA closes this gap: it is *patient-specific*, discharge-centred, and evaluates multi-label reasoning for three core tasks—diagnosis, medication, and instruction—under the same partial-information constraints clinicians face.
- We establish a time-aware evaluation protocol that prevents information leakage by enforcing chronological splits; all splits and data-preparation code are released to ensure reproducible, leakage-free evaluation.

The EXPRAG method.

- To our knowledge (Feb 2025 draft) we are the first to exploit *structured* EHR similarity

(ICD, procedure, and drug codes) to retrieve experience cases before applying text retrieval for discharge reasoning.

- The pipeline uses only tabular codes, making it substantially cheaper and more privacy-robust than systems that embed entire notes or build extensive ontologies. Recent work such as MedRAG (Zhao et al., 2025) and KARE (Jiang et al., 2025) focuses on knowledge-graph or multi-source approaches; by contrast, EXPRAG reuses routine ICD/NDC tables to surface "near-patient" experience without KG construction overhead.

- Controlled ablations confirm that adding these structured features alone delivers a statistically significant accuracy lift across all three tasks.

- At the framework level, EXPRAG departs from generic fact retrieval: it retrieves clinically similar past cases and uses structured EHR signals to gate the cohort prior to text retrieval. This case-level, EHR-grounded retrieval is complementary to KG-based or multi-source approaches and provides an auditable clinical prior that conditions downstream text retrieval.

An end-to-end, clinically grounded evaluation pipeline.

- We deliver a fully reproducible end-to-end pipeline and experiments for clinical discharge support, demonstrating that a coarse-to-fine design (structured EHR gating $\rightarrow$ text retrieval $\rightarrow$ LLM reasoning) improves downstream decision support when compared to baselines that use ungated dense retrieval.

- Empirically, the framework is consistently effective across multiple LLM backbones and text retrievers: EXPRAG improves end-task performance and the EHR-based ranker outperforms several embedding-based rankers within the same coarse-to-fine protocol.

- Gating with structured EHR signals avoids dense retrieval/reranking over the entire corpus, which reduces compute, token consumption, and latency—so a simpler (and practically cheaper) mechanism yields stronger empirical results and can be deployed as a drop-in component in larger systems.

- Simplicity is deliberate for auditability: transparent set-similarity (Jaccard with a weighted aggregation) makes every match traceable to concrete codes, which is essential for clinical review, safety, and hospital deployment.

Together, these contributions (dataset, method, and evaluation pipeline) provide a reproducible testbed and a practically deployable retrieval strategy for discharge-centered clinical reasoning.

A.10.2 Q2 — Why do we use multiple-choice questions rather than free-text?

Most electronic health-record (EHR) systems already present clinicians with *structured pick-lists*: diagnoses come from ICD or SNOMED menus, prescriptions from formulary look-ups, and discharge instructions from hospital-approved templates. Selecting the correct item from a list therefore mirrors everyday practice far better than asking clinicians to compose free-text from scratch.

For research, a multiple-choice (MCQ) design has two additional advantages:

- **Objective scoring.** MCQs enable automatic exact-match or F_1 grading without costly expert adjudication, keeping the benchmark reproducible.

- **Still challenging for LLMs.** Current state-of-the-art models score $\leq 64\,F_1$ on our tasks; thus the MCQ setting remains far from saturated.

Free-text evaluation is on our roadmap, but we prioritised MCQs for this release to balance clinical realism with reliable, low-variance scoring.

A.10.3 Q3 — Does the low score on *Medication* arise from LLM safety refusals?

No. A complete audit of the GPT-3.5 and GPT-4o outputs used in our benchmark revealed no refusals; more than 99 % of generations were valid answers with an accompanying rationale.

Why is the task still hard?

- **Wide outcome space.** Discharge drug regimens vary greatly across patients; small differences in comorbidities or procedures can shift the correct therapy.

- **Missing clinical context.** Safe prescribing hinges on laboratory trends, imaging, and procedural details that are *not* contained in the discharge note. Models must reason with these latent variables or accept higher uncertainty.

Adding labs and other modalities would almost certainly improve accuracy. Our code-based retriever provides a scalable path to incorporate these data, and the current results should be viewed as a conservative lower bound that future multimodal systems can surpass.

A.10.4 Q4 — What happens when the target patient has no close match in MIMIC-IV?

Our retrieval pipeline (Fig. 2, §3.3) is *coarse-to-fine*. A Jaccard filter first scores structured codes (ICD, procedures); missing codes simply contribute 0 to the numerator, so patients with partial tables remain comparable. Even without near duplicates, the filter still surfaces cases that share at least some comorbidities, ensuring the pool is never empty. If all structured scores fall below a threshold, the system widens the slice and falls back to pure text retrieval—no question is answered "from scratch".

Graceful degradation. In Table 5 ("Complementary" ablation) we *zero-weight the task-relevant modality*. The end-to-end F_1 drops by less than 3 %, confirming that the method remains robust when one or more code tables are sparse.

A.10.5 Q5 — Is there a chance of data leakage between the structured EHR tables and the discharge-note answers?

Distinct data sources. Structured tables (`diagnoses`, `medications`, `procedures`) capture *in-hospital* information, whereas gold answers are extracted from the clinician-authored *discharge summaries*. Many IV-only or acute-care drugs recorded in the Medication Administration Record (MAR) never appear in the take-home prescription list, underscoring the practical separation of the two sources.

Section masking eliminates label exposure. For every question we include only note sections that precede the clinical decision and *mask all later sections*—e.g. the discharge plan is hidden when we ask for a diagnosis—thereby preventing the model from "seeing the answer" in its context.

Controlled option generation. Correct answers come from the discharge note; distractors are sampled from the same patient's structured tables, and any accidental overlaps are automatically merged away. This guarantees that an option cannot be correct in both sources simultaneously.

Clinically faithful yet challenging. The workflow mirrors real practice: physicians distill noisy, high-volume in-hospital data into a concise discharge plan. The resulting multi-select questions compel models to synthesize information instead of memorising patterns, while remaining leakage-free.

A.10.6 Q6 — What practical steps could improve performance?

Enrich the evidence pool. Our current study constrains retrieval to *structured codes + discharge notes from about 400 k admissions*. Future users can plug additional EHR signals—laboratory trends, imaging reports, vital-sign time-series, even longitudinal clinic notes—into the same sparse-code index to approximate the full information set clinicians use. ("Limitations") lists the required table names and normalisation scripts.

Insights from initial failure analysis. We manually reviewed 40 failed cases using ExpRAG, comparing them with results from direct LLM responses, to identify common failure patterns.The three most common error categories identified were: (1) Incomplete retrieval (2) Partial information (3) Numerical inaccuracies.

Insights from futher failure analysis. We developed a pipeline using GPT-4o to systematically categorize each failure by providing the original question, answer options, and LLM rationale.

- **23 %** of failures were due to retrieved similar patients lacking all correct answer points.
- **47 %** involved cases where the correct information was present but incomplete in the retrieved content.
- **15 %** were attributed to numerical inaccuracies in the retrieved content (e.g., "2 weeks" vs. "5 days").

Take-away. Even within the current "codes + notes" boundary, improving retrieval completeness and numeric fidelity could close roughly half of the remaining error gap; adding labs and imaging should push performance further while keeping the workflow clinically plausible.

A.10.7 Q7 - How to reproduce EXPRAG on DISCHARGEQA?

Codes and sample dataset can be found from https://anonymous.4open.science/r/EXPRAG-3804

A.10.8 Q8 — If no new expert review was performed, how can the dataset still be trusted?

Clinician-authored ground truth. Every correct answer is copied verbatim from the discharge summary written by the patient's treating clinician in MIMIC-IV (see §4.1). Because those summaries constitute the hospital's official record, they already carry the clinical validation that external annotation would otherwise provide.

Use of GPT-4o is limited to distractors. Large-language models were employed *only* to craft plausible but incorrect answer options; they never determined which option is correct. This separation avoids the circularity that can arise when the same model both generates and grades answers.

Quality spot-checks. To confirm that the automated pipeline preserved accuracy, we randomly sampled 200 QA pairs. Two teamembers independently verified the matching of the labelled answers with original discharge notes with 100 % alignment.

Outcome. By leveraging the original clinician notes and restricting LLM use to distractor generation—supplemented by light expert spot-checking—we obtain a high-credibility dataset without incurring the heavy cost of full manual annotation.

A.10.9 Q9 — When and how will DISCHARGEQA be released?

Host. The corpus will be deposited on **PhysioNet** under the *credentialed-health-data* tier, matching MIMIC-IV's access model.

Eligibility. Anyone with an active MIMIC-IV Data-Use Agreement (DUA) and completed CITI "Data or Specimens Only" training can download the files immediately; new users follow the same two-step application.

Timeline. The DOI and metadata record will go live upon camera-ready acceptance. Until then, reviewers can inspect a pre-release snapshot in the supplementary material.

Code. All extraction, distractor-generation, and evaluation scripts will be open-sourced on GitHub under the MIT licence, synchronised with the PhysioNet release.

A.10.10 Q10 - Why not run cross-evaluation on EHRNoteQA/MedQA?

EXPRAG is designed for EHR-grounded, patient-centric QA, where the first-stage ranker depends on structured patient entities such as diagnoses, procedures, and medications. This requirement does not align cleanly with many existing clinical QA benchmarks. For example, MedQA is an exam-style knowledge QA dataset without patient cohorts or structured EHR tables, while note-centric datasets such as EHRNoteQA do not always expose the compatible tabular features required for our similarity stage.

For this reason, we do not view these benchmarks as apples-to-apples evaluation settings for the current EXPRAG pipeline. Applying EXPRAG to such datasets would require either reconstructing compatible structured patient representations or substantially modifying the retrieval pipeline, which is beyond the scope of this initial work.

At the same time, this also highlights an important limitation of the present study: our evidence is currently restricted to a single MIMIC-IV-derived cohort. Although this matches the intended deployment setting of within-institution EHR retrieval, broader validation across additional EHR datasets and multi-center databases remains important future work.

Scope & complementarity. Our work targets discharge-workflow reasoning (diagnoses, take-home medications, and instructions) rather than generic span extraction over full notes or exam-style QA. Table 1 contrasts DISCHARGEQA with prior corpora: EHRNoteQA/CliniQG4QA emphasize extraction from complete discharge notes, and MedQA/MedMCQA evaluate exam questions; DISCHARGEQA centers on discharge decisions under partial-information constraints. We will make this scope explicit and position our benchmark as complementary to prior datasets.

Prompt: generate options for Diagnosis task

Role: You are a doctor evaluating the Discharge Diagnosis of a patient.

Task: Your objective is to review the discharge diagnosis provided in the discharge summary and determine whether these diagnosis are suitable for the patient's treatment plan. The correct options are based on the diagnosis listed in the discharge summary, while incorrect options are derived from diagnoses table that are not part of the discharge diagnosis but may have been found during the hospital stay.

Discharge Diagnosis:
----info starts----
{discharge_diagnosis}
----info ends----

Diagnoses Database Info:
----info starts----
{diagnoses}
----info ends----

Also, please review the provided background info from other part of the Discharge Summary, which can be summarized (keep important info) to be background info, but do not put any diagnosis decision into it.
----background info starts----
{discharge_summary}
----background info ends----

Please provide a multi-answer true/false response for the following question:

Question:
Which discharge diagnosis were made for the patient at discharge?

Answer Options:
Provide a list of diagnosis in JSON format.
Each diagnosis should be marked as "True" if it was in the discharge diagnosis and "False" if it was only found in the diagnoses history but not listed as a discharge diagnosis.

Instruction:
1. List all items in the diagnosis and assign one option letter (from A to Z then a to z) to each non-repeated one
2. Review all items provided in the diagnoses database one by one. If the item is also listed by discharge diagnosis, or equivalent or very close meaning, then the "correct_answer" should have the letter of this item, and this item should be the same way as described in "Discharge Diagnosis"
3. If the item is from "Diagnoses Database Info" only but not in "Discharge Diagnosis", and the name is too long, please summarize it to be less than 10 words

Output Format:
Provide your responses in JSON format as follows:
{
"Reason": "<Explain how you combine equivalent diagnosis from both info sides to which options, and which options are from which info>"
"background": "{background} + <Your summary from other parts of the Discharge Summary, do not put diagnosis info into it, try to include as much important info as possible>",
"discharge_diagnosis_options": {
"A": "<diagnosis name>",
"B": "<diagnosis name>",
"C": "<diagnosis name>",
...
},
"correct_answer": "<String of options representing correct diagnosis, e.g., 'ACD'>"
}

Figure 7: Prompt design example for Diagnosis tasks

20453584

Background: a male physician with BUE paresthesias and periscapular pain, experiencing bilateral ulnar distribution symptoms, with MRI showing C5-6 disc-osteophyte complex and spinal canal narrowing.
"discharge_diagnosis_options": {
"A": "disc osteophyte",
"B": "Esophageal reflux",
"C": "Obstructive sleep apnea",
"D": "Diaphragmatic hernia",
"E": "History of skin neoplasm",
"F": "Brachial neuritis",
"G": "spinal stenosis"}
Correct Answer: A, G
Question: Which diagnoses should be documented into the patient's discharge diagnosis?

20453584

History of Present Illness:

...
This is an otherwise healthy male who complains over the past couple months of having **bilateral left greater than right ulnar distribution, paresthesias** as well as **periscapular pain**. MRI was most concerning for **C5-6: A disc-osteophyte complex with left sided herniation,** which causes **moderate spinal canal narrowing**. **Severe left and moderate right neural foraminal narrowing.**

Brief Hospital Course:
Patient was admitted to the ___ Spine Surgery Service and taken to the Operating Room for the above procedure. The surgery was **without complication** and the patient was transferred to the PACU in a stable condition. **TEDs/pnemoboots** were used for postoperative DVT prophylaxis. Intravenous antibiotics were continued for 24hrs postop per standard protocol. Initial postop pain was controlled with a **PCA**. Diet was advanced as tolerated. The patient **ambulated independently**. **Hospital course was otherwise unremarkable**. On the day of discharge, the patient was afebrile with stable vital signs, comfortable on oral pain control, and tolerating a regular diet.

Medications:
1. Patanol (olopatadine) 0.1 % ophthalmic BID
2. Ketorolac 0.5% Ophth Soln 1 DROP BOTH EYES QID dry eyes

Discharge Diagnosis:
disc osteophyte
spinal stenosis

28817667

History of Present Illness:

...
The patient has **neck pain and right arm pain** and **tingling**. He was found to have **C6-C7 disk herniation** with **right-sided radiculopathy**, which correlates with his symptoms.

Brief Hospital Course:
Patient was admitted to the ___ Spine Surgery Service and taken to the Operating Room for the above procedure. The surgery was **without complication** and the patient was transferred to the PACU in a stable condition. **TEDs/pnemoboots** were used for postoperative DVT prophylaxis. Intravenous antibiotics were continued for 24hrs postop per standard protocol. Initial postop pain was controlled with **oral pain medication**. Diet was advanced as tolerated. **The patient ambulated independently** post-op. **Hospital course was otherwise unremarkable**. On the day of discharge, the patient was afebrile with stable vital signs, comfortable on oral pain control, and tolerating a regular diet.

...
Medications:
1. **Gabapentin** 300 mg PO TID
2. **Oxycodone-Acetaminophen** (5mg-325mg) 1 TAB PO Q8H:PRN neck pain or dysmenorrhea
3. **Naproxen** 440 mg PO ONCE ___ TIMES PER WEEK

...
Discharge Diagnosis:
1. C6-C7 disk herniation.
2. C7 radiculopathy.

25633130

History of Present Illness:

...
This is a patient with a **large C3-C4 disc herniation above her previous construct** with **spinal cord changes**. Given her significant **upper extremity symptoms, myelomalacia** and **severe stenosis**, she is here for **surgical intervention**.

...

Discharge Diagnosis:
cervical stenosis with disc herniation

29378221

History of Present Illness:

...
Ms. ___ had previously undergone an **ACDF** with Dr. ___ at ___ in ___. When she was recovering from this surgery, she began having **low back pain** and **bilateral leg pain** down the posterolateral aspect of both thighs with some soreness in her calves. An MRI showed a **facet cyst at L4-L5**, and **she underwent cyst aspiration,** which relieved her pain until it returned. She does not report specific weakness in her legs.

...
Discharge Diagnosis:
L4/5 central stenosis/disc hernia

Figure 8: Comparison of Similar Patients.