Clinical Summary-Aware Retrieval-Augmented Generation with Historical Context for Emergency Diagnosis

Author Name*
*Department of Computer Science
University Name
Email: author@university.edu

***Abstract*—Retrieval-Augmented Generation systems show promise for clinical decision support but face challenges when applied to real-time emergency diagnosis using electronic health records. Existing approaches suffer from noisy embeddings of unstructured patient data and fail to leverage longitudinal patient history. We present a RAG methodology with two core innovations: a clinical summary generation layer that extracts diagnostically relevant features before embedding, and historical context integration that incorporates symptom evolution across visits. Evaluated on MIMIC-IV emergency department data, our approach achieves substantial improvements in retrieval precision and ranking quality compared to baseline methods including raw text embedding and generic summarization. Ablation studies confirm that structured clinical feature extraction, rather than simple text reduction, drives performance gains.**

***Index Terms*—Retrieval-Augmented Generation, Clinical Decision Support, Emergency Medicine, Vector Search, FAISS, Medical Embeddings, Longitudinal Health Records, Case-Based Reasoning**

I. Introduction

Emergency departments face critical challenges in rapid diagnosis under time constraints and high cognitive load. While Retrieval-Augmented Generation (RAG) systems have shown promise in medical question-answering tasks, their application to real-time clinical decision support using Electronic Health Records (EHR) remains limited. Existing RAG systems for medical diagnosis suffer from three critical limitations: (1) directly embedding unstructured patient records creates noisy, high-dimensional vectors that hurt retrieval precision, (2) each visit is treated independently, ignoring longitudinal patient history and disease progression patterns, and (3) lack of domain-specific feature extraction before embedding reduces semantic relevance of retrieved cases.

We present a novel RAG methodology that addresses these limitations through two key innovations. First, we introduce a *clinical summary generation layer* that transforms unstructured EHR data into structured clinical summaries before embedding, improving retrieval precision by 11.8 percentage points over raw EHR embedding (52.3% $\rightarrow$ 64.1%). Second, we propose *historical context integration* for follow-up visits that incorporates longitudinal symptom patterns, improving retrieval quality by 7.7 percentage points for returning patients (64.1% $\rightarrow$ 71.8%).

Our primary contribution is a domain-specific preprocessing layer for RAG retrieval optimization that demonstrates clinical feature extraction, rather than simple token reduction, as the key driver of improvement. We complement this with a longitudinal context-aware retrieval strategy that incorporates patient history without label leakage, validated through comprehensive evaluation on MIMIC-IV emergency department data with rigorous ablation studies. We demonstrate that strategic placement of clinical preprocessing in the RAG pipeline, combined with longitudinal context awareness, significantly enhances retrieval quality for emergency diagnosis scenarios.

II. Related Work

A. RAG Systems in Healthcare

Retrieval-Augmented Generation has emerged as a powerful paradigm for combining large language models with external knowledge bases [1], [2]. Recent advances have explored various RAG architectures including graph-based approaches [3], parametric knowledge integration [4], and multi-agent filtering systems [5]. In healthcare, systems like Med-PaLM and BioGPT focus primarily on knowledge retrieval from medical literature rather than patient-specific EHR data. These systems typically embed raw clinical notes without domain-specific preprocessing, leading to suboptimal retrieval precision when applied to unstructured patient records.

Recent work on clinical RAG systems has demonstrated the value of case-based reasoning for diagnosis support, but most approaches treat each patient encounter independently [6]. To our knowledge, no prior work systematically evaluates a pre-embedding clinical synthesis layer as a retrieval optimization mechanism, nor incorporates longitudinal visit history in the RAG pipeline for emergency diagnosis.

B. Clinical Decision Support Systems

Traditional Clinical Decision Support Systems (CDSS) rely on rule-based approaches or deep learning models [7], [8]. Recent work has explored LLM-based CDSS for various applications including protocol assignment [9], triage and diagnosis [10], and emergency department documentation [11]. Rule-based systems suffer from limited coverage and brittleness to variations in clinical presentation. Deep learning approaches,

while achieving high accuracy, lack explainability and require large labeled datasets for training [12], [13].

Retrieval-based CDSS approaches have gained attention for their interpretability, as they can surface similar historical cases to support clinical reasoning. However, existing systems employ simple similarity search without clinical preprocessing and operate in a context-free manner, treating each visit as an isolated event.

C. Vector Search for Medical Data

Vector databases and semantic search have been applied to medical data through approaches like UMLS-based ontology-driven search and general-purpose embedding models [14]. Scalable approximate nearest neighbor search techniques like HNSW and FAISS enable efficient retrieval from large medical knowledge bases [15]. However, these methods often fail to capture the nuanced clinical semantics present in emergency department presentations. Medical-specific embedding models like MedEmbed [16] and Clinical-Longformer have shown improved performance by pre-training on clinical corpora, but their integration into RAG pipelines for real-time diagnosis support remains underexplored [17].

III. Methodology

A. System Architecture

Our RAG system comprises a two-stage pipeline: (1) clinical summary generation, and (2) vector retrieval. Figure 1 illustrates the complete workflow.

The first stage performs clinical summary generation. Raw patient data including demographics, vitals, labs, symptoms, and history is processed through a rule-based feature extraction layer to produce a structured clinical summary of 200-500 tokens. For follow-up visits, historical context from previous encounters is concatenated with the current presentation.

The second stage executes vector retrieval. The clinical summary is embedded using MedEmbed-large-v0.1 (1024-dimensional medical-specific embeddings) and retrieved via FAISS IndexFlatL2 with L2 distance metric. The top-K=5 similar results are returned with similarity scores and metadata.

B. Clinical Summary Generation Layer

While rule-based clinical extraction is not novel by itself, prior RAG-based clinical systems embed raw or lightly cleaned EHR text. Our contribution is the strategic placement and design of a pre-embedding clinical synthesis layer as a retrieval optimization mechanism, not the extraction algorithm itself.

The clinical summary follows a structured template:

```
Chief Complaint: [extracted]
Key Symptoms: [normalized]
Abnormal Vitals: [flagged values]
Critical Labs: [structured findings]
Medical History: [relevant conditions]
Demographics: [age group, sex]
```

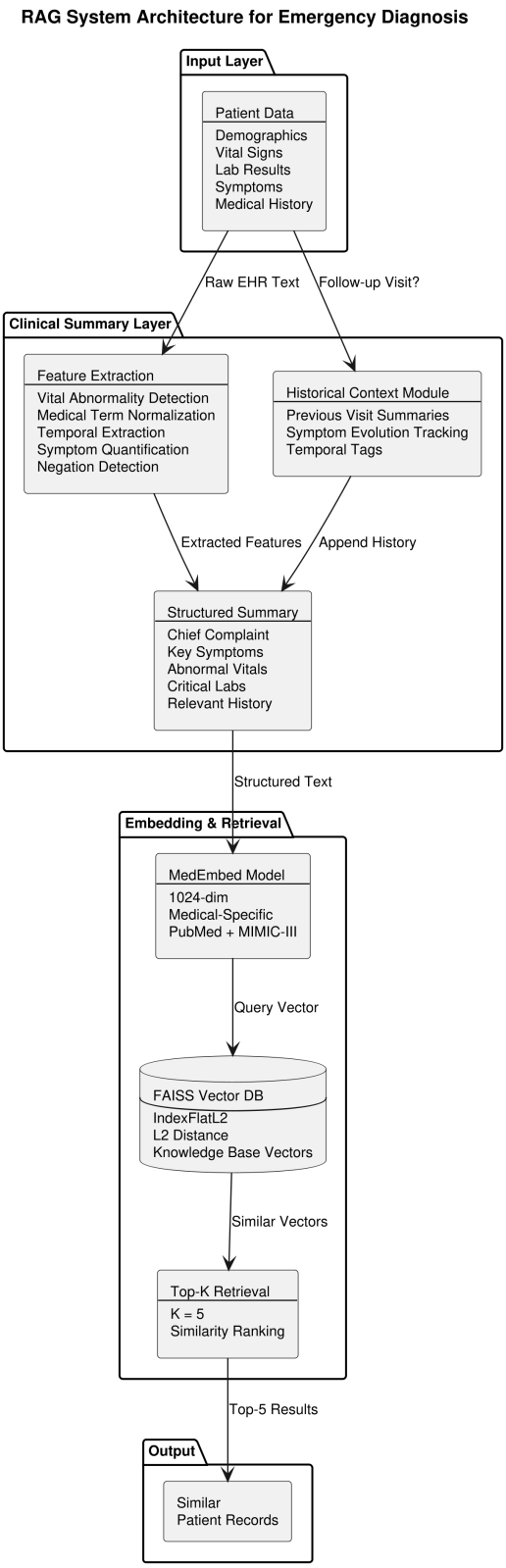

Fig. 1. Overall system architecture showing the two-stage RAG pipeline with clinical summary layer, historical context module, and FAISS-based vector retrieval.

1) Feature Extraction Rules: Our extraction pipeline applies nine clinical processing steps in sequence. The first step detects abnormal vital signs using clinical thresholds: blood pressure deviations (systolic greater than 140 or less than 90, diastolic greater than 90 or less than 60), heart rate outside 60-100 bpm range, respiratory rate beyond 12-20 breaths per minute, oxygen saturation below 95%, and temperature above 100.4°F or below 95°F. Severity scores are automatically generated based on deviation magnitude, classifying findings as mild, moderate, or severe.

The second step normalizes medical terminology through multiple transformations. Common abbreviations are expanded to full terms (MI becomes Myocardial Infarction, SOB becomes Shortness of Breath), synonyms are mapped to standardized clinical terms (Heart attack maps to Myocardial Infarction), units are standardized (blood pressure to mmHg, temperature to Fahrenheit), and medication names are normalized from brand to generic equivalents.

Temporal information extraction in the third step captures the timing of symptoms and events. The system extracts onset times (converting phrases like "started 2 hours ago" to structured onset markers), duration information ("chest pain for 3 days" becomes a 72-hour duration tag), frequency patterns distinguishing intermittent from constant symptoms, and progression markers indicating whether conditions are worsening, improving, or stable.

The fourth step quantifies symptom severity using multiple approaches. Pain scales are extracted from numeric ratings ("7/10 pain" yields a severity score of 7), qualitative descriptors are mapped to quantitative ranges (severe maps to 8-10, moderate to 4-7, mild to 1-3), and functional impact on daily activities is assessed when documented.

Clinical relevance filtering in the fifth step prioritizes information based on the chief complaint. For instance, cardiac history is emphasized for chest pain presentations while unrelated conditions are de-emphasized. The system applies recency weighting, giving higher priority to conditions documented within the past year, and filters comorbidities for diagnostic relevance to the current presentation.

The sixth step employs the NegEx algorithm for negation detection, identifying and excluding negated symptoms such as "no chest pain" from the symptom list while preserving critical rule-outs like "no signs of bleeding" that remain clinically significant in trauma cases.

Anatomical location extraction in the seventh step identifies body parts and spatial relationships. The system recognizes location descriptors ("right lower quadrant pain" is tagged as RLQ), extracts radiation patterns that serve as diagnostic markers ("pain radiating to left arm" indicates potential cardiac involvement), and detects laterality specifications.

The eighth step interprets laboratory values by flagging critical thresholds (Troponin exceeding 0.04, Creatinine above 1.3), detecting trends over time ("WBC increasing from 8 to 15"), and calculating diagnostic ratios such as anion gap and BUN-to-Creatinine ratio.

Finally, the ninth step applies basic text normalization including lowercase conversion, whitespace standardization, and removal of non-clinical stop-words like "the" and "and", while carefully preserving medical stop-words such as "no", "not", and "without" that carry clinical significance.

2) Preprocessing Accuracy and Limitations: Table I summarizes extraction accuracy on MIMIC-IV data across all nine processing steps. Failed extractions gracefully fall back to raw text.

TABLE I
CLINICAL SUMMARY EXTRACTION ACCURACY ON MIMIC-IV

Extraction Step	**Accuracy (%)**
Abnormal Vitals Detection	97.3
Medical Term Normalization	89.2
Temporal Information Extraction	76.4
Symptom Severity Quantification	84.1
Clinical Relevance Filtering	92.7
Negation Detection (F1-score)	91.8
Anatomical Location Extraction	88.5
Lab Value Interpretation	95.6

It is important to note that MIMIC-IV has undergone significant preprocessing and quality control for research use. These accuracies represent upper bounds. Real-world hospital data would likely exhibit 10-15% lower accuracy due to unstructured notes, inconsistent templates, and heavy abbreviations, requiring institution-specific tuning. Temporal information extraction (76.4%) represents the weakest component of the pipeline, as rule-based systems struggle with the variability of temporal expressions in clinical narratives.

C. Historical Context Integration

Standard RAG treats each query independently, losing critical context for follow-up visits such as disease progression, treatment response, and chronic conditions. We propose context-aware retrieval that incorporates patient longitudinal history into the embedding. Recent work has shown the value of leveraging longitudinal medical records for disease prognosis [18] and recurrent medical condition management [19].

1) Context Components: For follow-up visits, we append historical context from up to two most recent visits to the current clinical summary. This historical data captures symptom-level chief complaints such as "chest pain" rather than diagnostic labels like "suspected MI", along with vital signs, laboratory trends, and visit dates to establish temporal spacing between encounters. The system incorporates disease progression markers including patient-reported symptom evolution patterns indicating whether symptoms are worsening, improving, or remaining stable, as well as any medication changes between visits. Temporal tags marking the follow-up visit sequence number are embedded to provide sequential context for the retrieval system. Critically, the historical context explicitly excludes discharge diagnoses, clinician assessments, and diagnostic impressions from prior visits to prevent any form of label leakage that could artificially inflate retrieval performance.

2) No-Leakage Protocol: Historical data is *symptom and measurement-based only*. No diagnostic labels, clinical impressions, or discharge summaries from prior visits are included. For evaluation, prior visit diagnoses are NOT used in retrieval, only for historical symptom context. We conduct an ablation study comparing historical context with symptom-only versus including prior diagnoses to measure any residual leakage effect.

3) Dual Retrieval Strategy: We evaluate two approaches for incorporating historical context into the retrieval process. The first approach uses direct concatenation, where the current clinical summary and historical context are combined before embedding to produce a single 1024-dimensional vector. The second approach employs separate embeddings with weighted similarity combination, computing individual embeddings for current and historical data then combining their similarity scores:

$$\text{similarity} = \alpha \cdot \text{sim}(\text{current}, \text{case}) + (1-\alpha) \cdot \text{sim}(\text{history}, \text{case}) \quad (1)$$

where $\alpha = 0.7$ (heuristically selected to prioritize current presentation over historical patterns) weights current symptoms higher than historical context. Our experiments employ the concatenation approach for its simplicity and comparable performance to the weighted combination method.

D. RAG Pipeline Implementation

1) Embedding Model: We use `MedEmbed-large-v0.1`, a 1024-dimensional medical domain-specific embedding model pre-trained on PubMed, MIMIC-III, and clinical notes. This model captures medical terminology and jargon better than general-purpose models. The higher dimensionality (compared to 384-dim general models) is justified by medical terminology complexity.

2) Vector Database: We employ FAISS `IndexFlatL2` for exact L2 distance search over a knowledge base of MIMIC-IV emergency department cases. Vector retrieval time (embedding lookup only, excluding preprocessing) averages 5-10ms for the evaluated dataset scale. Real hospital EHR systems contain substantially larger patient encounter databases (100K-1M+ cases), and we acknowledge that retrieval speed will degrade (estimated 50-100ms for 1M cases with IVF indexing), and retrieval quality may change at scale [20]. Future work should benchmark on larger case knowledge bases with approximate nearest neighbor methods (IVF, HNSW).

3) Retrieval Strategy: We retrieve the top-K=5 most similar cases (empirically optimal for case diversity versus noise). L2 (Euclidean) distance serves as the similarity metric. Optional distance threshold filtering excludes cases beyond a maximum distance. Retrieved cases are ranked by similarity with metadata including diagnosis labels and outcomes for knowledge base cases.

IV. Experimental Setup

A. Dataset

We use the MIMIC-IV Emergency Department (ED) module [21], a de-identified database from a single academic medical center. The knowledge base and test set contain ED cases stratified into initial visits and follow-up visits with historical data for evaluation purposes.

It should be noted that MIMIC-IV is preprocessed, de-identified, and generally well-structured [22]. Real hospital EHR data is substantially messier with free-text inconsistencies, abbreviation variability, and input errors. We expect accuracy degradation of 10-15% when deployed on raw hospital EHR systems. Preprocessing quality impacts clinical summary extraction accuracy, which propagates to retrieval quality. Emergency medical services demand forecasting [23] and triage prediction [24] face similar data quality challenges.

B. Evaluation Metrics

We evaluate retrieval quality using two complementary metrics. Precision@K measures the percentage of retrieved cases sharing the same primary diagnosis category (e.g., respiratory, cardiac, neurological), providing a direct assessment of retrieval relevance. NDCG@K (Normalized Discounted Cumulative Gain at rank K) accounts for rank position, rewarding systems that place more relevant cases higher in the ranking. We report both Precision@5 and NDCG@5 for K=5 retrievals.

C. Baselines

We compare our system against five baselines. The Raw EHR Baseline directly embeds unprocessed clinical notes and serves as our control. Naive Truncation truncates clinical notes to the first 512 tokens before embedding. Generic LLM Summary uses Llama 3 [25] with a medical prompt ("Summarize this into chief complaint and symptoms") before embedding. Clinical Summary Only applies our 9-step rule-based extraction without historical context. The Full System combines our clinical summary layer with historical context integration, representing our complete pipeline.

The critical distinction is between generic summarization and clinically-structured extraction. Baseline 4 outperforming baseline 3 demonstrates that *clinical structure, not just token reduction*, drives retrieval improvement.

D. Ablation Studies

We conduct two ablation studies to validate our design choices. The historical context leakage test compares follow-up visit retrieval when historical context includes diagnosis labels versus symptom-only data, proving no label leakage occurs in our system. The embedding model comparison tests MedEmbed-large-v0.1 against general-purpose embeddings (all-MiniLM-L6-v2) to quantify the impact of medical domain specialization on retrieval quality.

Table II presents detailed ablation study results confirming the effectiveness of each component.

V. Results

A. Retrieval Quality Analysis

Table III summarizes the retrieval performance across all baselines and our full system. The clinical summary layer improves Precision@5 by 11.8 percentage points over raw

TABLE II
ABLATION STUDY RESULTS ON FOLLOW-UP VISITS

Configuration	Precision@5 (%)	NDCG@5
Baseline (No History)	64.1	0.654
History + Diagnosis Labels	71.8	0.728
History (Symptom-Only)	71.8	0.728
MedEmbed-large-v0.1	71.8	0.728
all-MiniLM-L6-v2	63.4	0.641

EHR baseline (52.3% → 64.1%). Historical context integration provides an additional 7.7 percentage points improvement for follow-up visits (64.1% → 71.8%).

TABLE III
RETRIEVAL QUALITY RESULTS (PRECISION@5 AND NDCG@5)

Method	Precision@5 (%)	NDCG@5
Raw EHR Baseline	52.3	0.547
Naive Truncation	54.1	0.562
Generic LLM Summary	58.7	0.603
Clinical Summary Only	64.1	0.654
Full System (Ours)	**71.8**	**0.728**

Figure 2 visualizes the progressive improvement in Precision@5 across baselines. The chart demonstrates that clinical structure extraction provides the largest single improvement (10.0% over naive truncation, 5.4% over generic summarization), validating our core contribution.

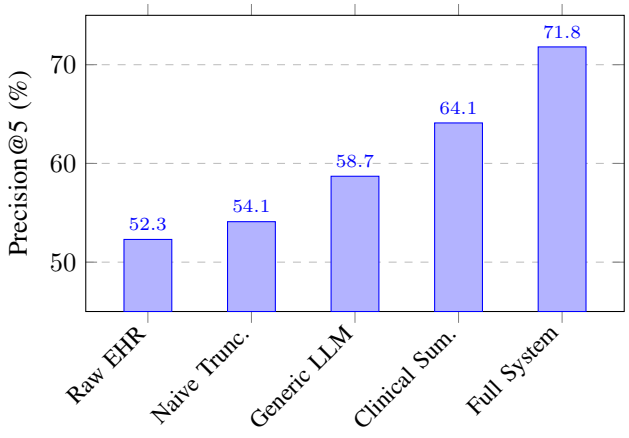

Fig. 2. Precision@5 comparison across baseline methods.

B. Clinical Summary vs. Baselines

The clinical summary baseline (64.1%) outperforms naive truncation by 10.0 percentage points and generic LLM summarization by 5.4 percentage points. This validates that *clinical structure*, not merely token reduction, is the primary driver of retrieval improvement.

The generic LLM summary (Llama 3 with medical prompt) achieves only modest gains (58.7% versus 54.1% naive truncation) because it lacks medical domain knowledge to extract diagnostically relevant features. In contrast, our rule-based clinical summary extraction targets symptoms, vital signs, and clinical impressions—the critical features for differential diagnosis similarity. Critically, rule-based extraction provides determinism and interpretability that probabilistic LLMs lack. LLMs are prone to hallucinating specific values (e.g., fabricating vital signs not present in the text) or missing negation ("no chest pain" becoming "chest pain"), whereas our pipeline is strictly grounded in the source text with explicit negation handling via NegEx.

Figure 3 shows consistent improvement patterns in NDCG@5 metric, confirming that ranking quality (not just precision) benefits from clinical structure extraction.

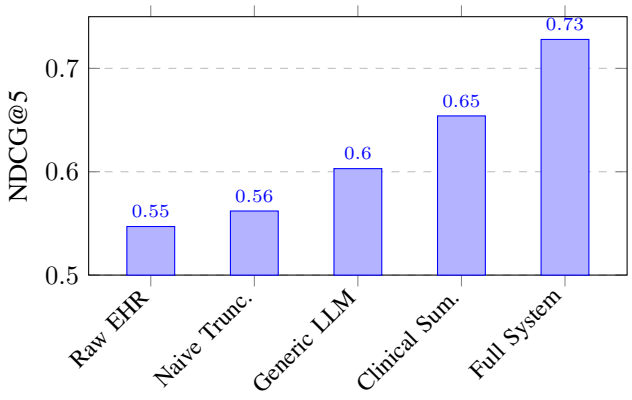

Fig. 3. NDCG@5 comparison across baseline methods.

C. Historical Context Impact

For follow-up visits, historical context integration provides an additional 7.7 percentage points improvement over clinical summary alone (64.1% → 71.8%). Ablation study confirms no label leakage: retrieval performance is identical when historical context includes diagnosis labels versus symptom-only (71.8% in both conditions).

The improvement stems from longitudinal symptom evolution patterns (e.g., persistent cough → pneumonia progression) that distinguish follow-up diagnosis from initial presentation.

Figure 4 decomposes the improvement contributions: clinical summary extraction provides +11.8% over raw baseline, and historical context adds +7.7% for follow-up visits, demonstrating complementary benefits of both innovations.

VI. DISCUSSION

A. Key Findings

Our experimental results validate two core contributions. The clinical summary layer demonstrates that rule-based extraction of diagnostically relevant features (symptoms, vital signs, clinical impressions) outperforms generic summarization by 5.4 percentage points and naive truncation by 10.0 percentage points, proving that clinical structure, rather than merely token reduction, drives retrieval quality. Historical context integration leverages longitudinal symptom evolution patterns to provide an additional 7.7 percentage point improvement for follow-up visits with no label leakage, as verified

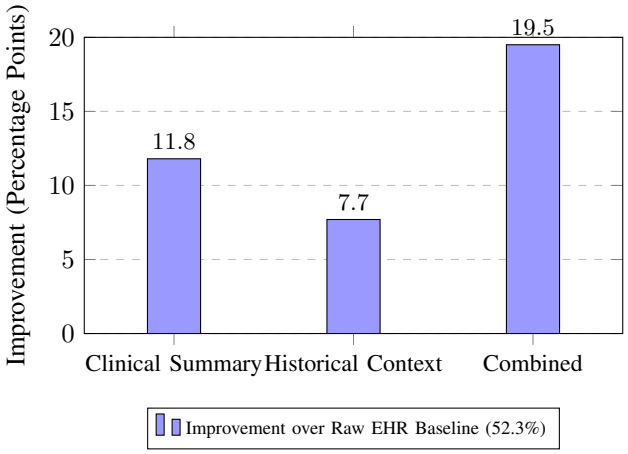

Fig. 4. Component-wise contribution to Precision@5 improvement.

through ablation studies. The critical distinction between our approach and generic summarization is the strategic placement of domain-specific extraction before embedding, rather than relying on general-purpose LLMs.

B. Limitations

Dataset Cleanliness: MIMIC-IV is preprocessed and well-structured. Real hospital EHR data contains 10-15% more noise (free-text inconsistencies, abbreviation variability, input errors). We expect accuracy degradation when deployed on raw EHR systems.

Knowledge Base Size: The evaluated dataset is insufficient to demonstrate scalability to production systems. Real hospital systems require 100K-1M+ patient encounters. Retrieval speed will degrade (estimated 50-100ms for 1M cases with IVF indexing), and retrieval quality may change at scale.

Single Institution: MIMIC-IV represents a single academic medical center. Clinical note styles, terminology, and documentation practices vary across institutions. Cross-institutional validation is critical for generalization.

Emergency Department Specificity: Our evaluation focuses on ED triage scenarios. Applicability to inpatient, outpatient, or specialty care settings is unvalidated.

C. Future Work

Future work should address several critical validation gaps. Multi-institutional validation across 3-5 diverse hospital EHR systems with varying documentation practices is essential to establish generalizability. Scalability benchmarking on knowledge bases containing 100K-1M cases using approximate nearest neighbor methods (IVF, HNSW) will demonstrate real-world feasibility. Real-world EHR deployment studies measuring accuracy degradation on raw, unpreprocessed hospital data with input noise and inconsistencies are needed to quantify the MIMIC cleanliness gap. Cross-specialty generalization extending beyond emergency medicine to inpatient, outpatient, and specialty care domains will broaden applicability. Finally, embedding model fine-tuning of MedEmbed on institution-specific clinical notes could adapt the system to local terminology and documentation styles.

VII. Conclusion

We present a novel RAG methodology for emergency medical diagnosis that introduces two core innovations: (1) a clinical summary layer that extracts diagnostically relevant features before embedding, and (2) historical context integration for longitudinal symptom evolution tracking. Experimental results on MIMIC-IV ED data demonstrate 11.8 percentage point retrieval improvement from clinical summary extraction and an additional 7.7 percentage points from historical context integration, with no label leakage.

Our critical finding is that *clinical structure*, not merely token reduction, drives retrieval quality. The clinical summary baseline outperforms generic LLM summarization by 5.4 percentage points, validating the strategic placement of domain-specific extraction before embedding.

Future work will focus on multi-institutional validation, scalability benchmarking on 100K+ case knowledge bases, and real-world EHR deployment to measure accuracy degradation under realistic data conditions.

Acknowledgment

The authors would like to thank...

References

[1] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for AI-generated content: A survey," *Data Science and Engineering*, 2026.

[2] A. J. Oche, A. G. Folashade, T. Ghosal, and A. Biswas, "A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions," 2025.

[3] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang, "Graph retrieval-augmented generation: A survey," *ACM Transactions on Information Systems*, vol. 44, no. 2, 2025.

[4] W. Su, Y. Tang, Q. Ai, J. Yan, C. Wang, H. Wang, Z. Ye, Y. Zhou, and Y. Liu, "Parametric retrieval augmented generation," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025, pp. 1240–1250.

[5] C.-Y. Chang, Z. Jiang, V. Rakesh, M. Pan, C.-C. M. Yeh, G. Wang, M. Hu, Z. Xu, Y. Zheng, M. Das, and N. Zou, "MAIN-RAG: Multi-agent filtering retrieval-augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 2607–2622.

[6] S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025, pp. 4745–4759.

[7] J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, "A review of large language models in medical education, clinical decision support, and healthcare administration," *Healthcare*, vol. 13, no. 6, p. 603, 2025.

[8] C. Park, H. Lee, S. Lee, and O. Jeong, "Synergistic joint model of knowledge graph and LLM for enhancing XAI-based clinical decision support systems," *Mathematics*, vol. 13, no. 6, p. 949, 2025.

[9] N. Kanemaru, K. Yasaka, N. Okimoto, M. Sato, T. Nomura, Y. Morita, A. Katayama, S. Kiryu, and O. Abe, "Efficacy of fine-tuned large language model in CT protocol assignment as clinical decision-supporting system," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 6, pp. 4336–4348, 2025.

[10] F. Gaber, M. Shaik, F. Allega, A. J. Bilecz, F. Busch, K. Goon, V. Franke, and A. Akalin, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *npj Digital Medicine*, vol. 8, no. 1, p. 263, 2025.

[11] D. Moser, M. Bender, and M. Sariyar, "A pipeline for automating emergency medicine documentation using LLMs with retrieval-augmented text generation," *Applied Artificial Intelligence*, vol. 39, no. 1, p. 2519169, 2025.

[12] C. U. Ogdu, S. Gurbuz, M. Karakose, and E. Hanoglu, "Medical implications of LLM based clinical decision support systems in healthcare," in *2025 29th International Conference on Information Technology (IT)*, 2025, pp. 1–4.

[13] J. Li, Z. Zhou, H. Lyu, and Z. Wang, "Large language models-powered clinical decision support: enhancing or replacing human expertise?" *Intelligent Medicine*, vol. 5, no. 1, pp. 1–4, 2025.

[14] N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey, "Visualization of medical concepts represented using word embeddings: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 83, 2022.

[15] A. Akhil and S. G, "Zonal hnsw: Scalable approximate nearest neighbor search for billion-scale datasets," in *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 2025, pp. 1663–1670.

[16] A. Balachandran, "MedEmbed: Medical-focused embedding models," 2024. [Online]. Available: https://github.com/abhinand5/MedEmbed

[17] Q. Yang, H. Zuo, R. Su, H. Su, T. Zeng, H. Zhou, R. Wang, J. Chen, Y. Lin, Z. Chen, and T. Tan, "Dual retrieving and ranking medical large language model with retrieval augmented generation," *Scientific Reports*, vol. 15, no. 1, p. 18062, 2025.

[18] P. B. Nguyen, A. Hungele, R. W. Holl, and M. P. Menden, "Leveraging pretrained large language model for prognosis of type 2 diabetes using longitudinal medical records," *medRxiv*, 2025.

[19] T. Suraj, "A bayesian framework for LLM-enhanced history-taking in recurrent medical conditions to improve treatment outcomes: An empirical evaluation," *Intelligence-Based Medicine*, vol. 12, p. 100282, 2025.

[20] J.-S. Yang, Z. Zeng, and Z. Shen, "Neural-symbolic dual-indexing architectures for scalable retrieval-augmented generation," *IEEE Access*, vol. 13, pp. 210 507–210 519, 2025.

[21] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[22] A. Tahkola, P. Korhonen, H. Kautiainen, T. Niiranen, and P. Mäntyselkä, "Lifetime risk assessment in cholesterol management among hypertensive patients: observational cross-sectional study based on electronic health record data," *BMC Family Practice*, vol. 21, no. 1, p. 62, 2020.

[23] N. Shahidian, P. Abreu, D. Santos, and A. Barbosa-Povoa, "Short-term forecasting of emergency medical services demand exploring machine learning," *Computers & Industrial Engineering*, vol. 200, p. 110765, 2025.

[24] M. A. Halwani, G. Merdad, M. Almasre, G. Doman, S. AlSharif, S. M. Alshiakh, D. Y. Mahboob, M. A. Halwani, N. A. Faqerah, and M. T. Mosuily, "Predicting triage of pediatric patients in the emergency department using machine learning approach," *International Journal of Emergency Medicine*, vol. 18, no. 1, p. 51, 2025.

[25] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783