




# No cloud, no problem: secure and explainable offline AI agents for clinical Q&A with lightweight LLMs

Ali Salman<sup>1</sup>, Giuseppe Fico<sup>2</sup> and Ernesto Iadanza<sup>1,3</sup>

<sup>1</sup> Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy.

<sup>2</sup> Life Supporting Technologies, Universidad Politécnica de Madrid, 28040 Madrid, Spain.

<sup>3</sup> Council of Societies, International Federation for Medical and Biological Engineering, Paris, France.

`a.salman@student.unisi.it`

`gfico@lst.tfo.upm.es`

`ernesto.iadanza@unisi.it`

**Abstract.** The advancement of artificial intelligence (AI)-driven clinical decision support systems has improved healthcare automation, yet most rely on cloud-based models, raising concerns about data privacy, latency, and accessibility in resource-constrained settings. We present a fully offline AI-powered patient-level question-and-answer (Q&A) system that integrates modular agents and retrieval augmented generation (RAG) with on-premise large language models (LLMs) for clinical summarization and diagnostic support, built on the Medical Information Mart for Intensive Care (MIMIC-IV v3.1) dataset. We preprocess 6,365,019 intensive care unit (ICU) admission summaries into 384-dimensional embeddings using the all-MiniLM-L6-v2 Sentence-Transformer and index them with facebook AI similarity search (FAISS) in a hierarchical navigable small world flat (HNSWFlat) structure (M=32, efConstruction=200) at roughly 145 records/s to enable sub-second nearest-neighbor retrieval. A Retrieval Agent fetches relevant patient history, vital signs, and laboratory results; a Summarization Agent (Mistral) converts structured data into concise, coherent narratives; and a Diagnosis Agent (Gemma) proposes likely conditions from those narratives. We record clinician-provided gold-standard notes alongside each model response and employ a hybrid feedback scorer—combining expert-weighted clinical keywords with international classification of diseases (ICD) keyword matches—to refine outputs (scores capped at 1.0). In a held-out evaluation on 24 real-world clinical queries, our offline framework achieves competitive retrieval precision, recall-oriented understudy for gisting evaluation (ROUGE-L) and bidirectional encoder representations from transformers score (BERTScore) summary quality, and top-3 diagnostic accuracy—while guaranteeing full on-site data residency and zero reliance on external application programming interfaces (APIs). This work demonstrates that sub-8 billion-parameter LLMs can deliver secure, explainable clinical Q&A entirely offline, paving the way for deployment in privacy-sensitive or connectivity-limited healthcare environ-

ments. Future work will explore fine-tuning on medical corpora and dynamic agent orchestration for real-time hospital integration.

**Keywords:** Clinical decision support, Retrieval-augmented generation, On-premise language models, Explainable AI, Data privacy

## 1 Introduction

The rapid emergence of large language models (LLMs) has revolutionized natural language processing by enabling sophisticated text understanding and generation. In clinical contexts, LLMs have achieved near-expert-level accuracy on tasks such as note summarization, question answering, and decision support [1]. However, the vast majority of these systems depend on cloud-based application programming interfaces (APIs), which introduce significant challenges related to patient data privacy, regulatory compliance, such as health insurance portability and accountability act (HIPAA), network latency, and service availability, especially in resource-constrained or offline settings.

Retrieval augmented generation (RAG) mitigates some of these issues by combining dense vector retrieval over domain-specific corpora with LLM-based reasoning. Grounding generation in retrieved clinical records reduces the risk of hallucinations and improves factual accuracy. Prior works such as Almanac [2] and self-reflective RAG architectures demonstrate that offline RAG pipelines can perform well on clinical question-answering benchmarks without relying on cloud services. Yet these solutions often assume the availability of high-capacity hardware or monolithic on-premise servers, limiting their applicability on edge devices or standard clinical workstations.

To address these limitations, we introduce a fully offline, multi-agent framework for patient-level question answering built on the Medical Information Mart for Intensive Care (MIMIC-IV v3.1) dataset [3,4,5]. Our pipeline leverages lightweight LLMs—Mistral for summarization and Gemma for diagnosis—coupled with a facebook AI similarity search (FAISS)-backed vector store of 384-dimensional embeddings extracted from all-MiniLM-L6-v2. This architecture runs efficiently on commodity hardware (e.g., Apple M4 Max), ensuring sub-second retrieval times and strict on-site data governance. The system comprises three specialized agents: a Retrieval Agent that fetches the top- $k$  similar “patient summaries,” a Summarization Agent that generates concise clinical narratives, and a Diagnosis Agent that proposes likely conditions. We also log clinician-provided gold-standard notes alongside each response and apply a hybrid feedback scorer—combining expert-weighted clinical keywords with international classification of diseases (ICD) keyword matches—to refine outputs (scores capped at 1.0).

In a held-out evaluation on 24 real-world clinical queries, our offline framework achieves competitive performance in top- $k$  retrieval (Recall@1 = 0.875, Recall@3 = 0.875), summary quality (ROUGE-L = 0.48, BERTScore F1 = 0.82)<sup>4</sup>,

<sup>4</sup> Recall-oriented understudy for gisting evaluation (ROUGE-L); bidirectional encoder representations from transformers score (BERTScore)

and top-3 diagnostic accuracy, all while preserving full data residency and zero reliance on external APIs. By combining privacy-preserving design principles [6] with modular, lightweight models, we demonstrate that sub-8 billion-parameter LLMs can deliver secure, explainable clinical Q&A entirely offline, paving the way for deployment in privacy-sensitive or connectivity-limited healthcare environments. Future work will explore fine-tuning on medical corpora and dynamic agent orchestration for real-time hospital integration.

## 2 Related Work

Several recent studies have explored RAG frameworks tailored to medical Q&A and clinical decision support. Yang et al. propose a dual retrieval-and-ranking RAG pipeline combining dense embeddings with Elasticsearch to improve accuracy and responsiveness in medical text generation, demonstrating its efficacy on specialized healthcare datasets [7]. Zhang et al. leverage long-context transformers in a two-stage RAG system over ICU records, achieving marked improvements in retrieval fidelity and end-to-end Q&A performance on critical-care corpora [8].

Factuality-aware retrieval has been advanced by Kim et al. in their recurrence generation augmented retrieval (RGAR) framework, which integrates electronic health records (EHRs) and external corpora in a bidirectional retrieval-generation loop, obtaining state-of-the-art results on multiple clinical benchmarks with an 8 billion-parameter LLM [9]. Complementing these efforts, Ke et al. report on a Retrieval-Augmented Generation pipeline applied to 35 preoperative guidelines, demonstrating non-inferior accuracy (91.4% with RAG vs. 86.3% by junior doctors) and sub-20 s response times compared to 10 min human benchmarks, underscoring the need for rigorous safety and consistency evaluation in high-stakes clinical deployments [10].

Lee et al. present a benchmark at the AAAI Symposium on AI for Health of resource-efficient LLMs for edge deployment, showing that sub-8 billion-parameter models can deliver clinically relevant reasoning under strict latency and memory constraints [11]. Together, these works inform our fully offline, modular agent architecture that combines a FAISS-backed vector store with lightweight LLMs to enable secure, explainable clinical QA without cloud dependencies.

## 3 Materials and Methods

### 3.1 Dataset

We obtained credentialed access to the MIMIC-IV v3.1 dataset [3,4,5] and extracted the patients, admissions, diagnoses\_icd, chartevents, and labevents tables. We joined these tables on subject\_id and hadm\_id to reconstruct each intensive care unit (ICU) stay. To ensure data completeness, we retained only those stays for which demographic fields, at least one vital-sign measurement, and at least one laboratory result were available. All timestamps were normalized to

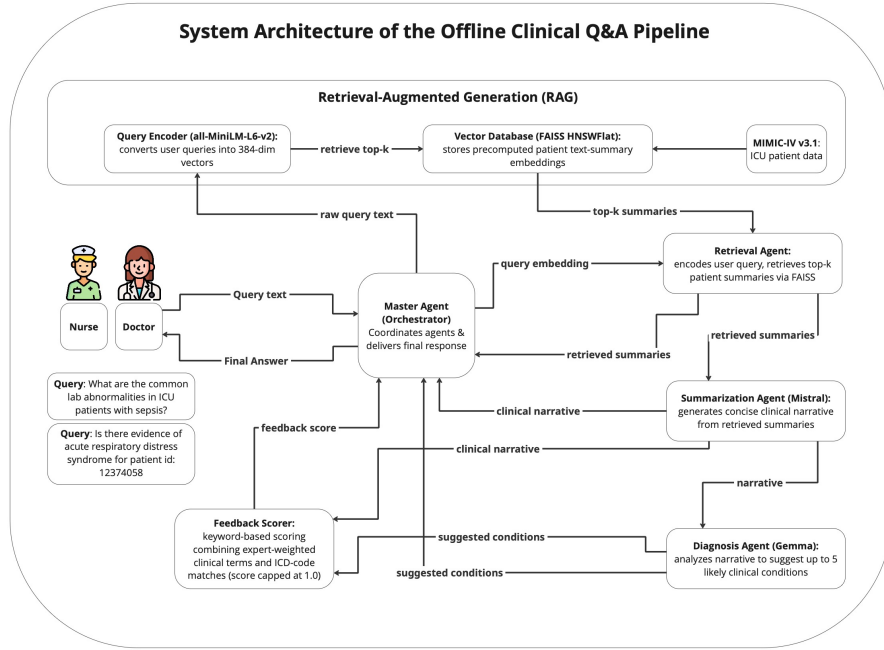


Fig. 1: An offline AI-driven clinical Q&A system leveraging AI agents, RAG, and lightweight LLMs for secure and explainable decision support. Designed for real-time retrieval, summarization, and diagnosis using MIMIC-IV v3.1 dataset.

ISO-8601 format, and measurement units were standardized (e.g., millimeters of mercury (mmHg) for blood pressure and beats per minute (bpm) for heart rate).

### 3.2 Summary Construction and Preprocessing

For each ICU admission, we constructed a single, pipe-delimited text summary containing patient demographics (subject ID, age, gender), admission meta-data (type, ISO-8601-normalized admit and discharge timestamps), the first ICD diagnosis code and its version, available admission vitals (heart rate in bpm; systolic/diastolic blood pressure in mmHg; SpO<sub>2</sub> as percent), and select laboratory values (creatinine, lactate). We then tokenized each summary with a 30,000-token SentencePiece model, truncated to 512 tokens, and wrote out text\_summaries.csv.gz with columns subject\_id, hadm\_id, and text\_summary for downstream embedding and FAISS indexing.

### 3.3 Embedding Generation and Vector Store

We encoded each text summary into a 384-dimensional float32 vector using the all-MiniLM-L6-v2 Sentence-Transformer in batches of 1024. These embeddings

were stored in a FAISS index using a hierarchical navigable small world flat (HNSWFlat) graph (parameters:  $M=32$ ,  $efConstruction=200$ ), co-located with the original summaries on disk to support sub-second k-nearest-neighbor retrieval.

### 3.4 Agent Pipeline

Our framework runs entirely offline on macOS with an Apple M4 Max (14-core CPU, 32-core GPU, 36 GB RAM). A single Master Agent orchestrates three specialized modules in sequence:

First, the **Retrieval Agent** converts a free-text user query into a 384-dimensional embedding using the all-MiniLM-L6-v2 encoder, then fetches the top- $k$  most similar “patient summaries” from the FAISS HNSWFlat index. Next, the **Summarization Agent** calls the Mistral LLM via Ollama’s CLI (temperature = 0.7, max\_tokens = 512), passing a system prompt together with those retrieved summaries to generate a concise, coherent clinical narrative. Finally, the **Diagnosis Agent** submits that narrative to the Gemma LLM (temperature = 0.8, max\_tokens = 256), requesting up to five likely clinical conditions.

Once all three agents have run, the Master Agent applies a keyword-based feedback scorer. Each candidate narrative  $D$  is assigned a score as in the equation below:

$$S = \sum_{i=1}^K w_i \mathbf{1}\{kw_i \in D\} + 0.01 H \quad (1)$$

where  $w_i$  is an expert-assigned weight for clinical keyword  $kw_i$  (e.g. “acute myocardial infarction”, “sepsis”, “ARDS”), the second term awards up to 0.1 points for distinct ICD-keyword matches, and  $H$  is the normalized Shannon entropy of the LLM’s token distribution. We then clamp  $S$  to the interval  $[0.1, 1.0]$ :

$$S \leftarrow \begin{cases} 0.1, & \text{if } S < 0.1, \\ 1.0, & \text{if } S > 1.0, \\ S, & \text{otherwise.} \end{cases}$$

Finally, the narrative with the highest  $S$  is selected as the final output. The full set of weighted keywords is listed in our supplementary code (scoring\_utils.py).

### 3.5 Baseline Comparisons and Evaluation

We evaluated our offline pipeline against two benchmarks. The first uses a cloud-based RAG model (facebook/rag-sequence-nq); the second follows a traditional information retrieval (IR) + LLM approach, retrieving documents with Okapi BM25 and summarizing them with GPT-2 (offline). Okapi BM25 scores each document by term frequency (TF), inverse document frequency (IDF), and document length normalization, ranking passages most relevant to the query.

We compared generated summaries both to held-out clinical notes (the “gold” MIMIC clinician summaries) and to our clinician\_note field extracted from the record, using ROUGE-L and BERTScore. We also measured diagnostic accuracy

(top-1 and top-3) against the ground-truth ICD codes and computed our automated `feedback_score`’s correlation with human quality judgments on a 24-query test set. Retrieval performance was reported as  $\text{precision@}k$  for  $k \in \{1, 3, 5\}$ , and we recorded end-to-end latency as well as peak memory usage on both CPU and MPS GPU backends. Finally, a qualitative review of fifty failure cases informed prompt refinements and final pipeline tuning.

Our full implementation is available on GitHub.<sup>5</sup>

## 4 Results

### 4.1 Retrieval Accuracy

We assessed our FAISS-based retriever—built on all-MiniLM-L6-v2 embeddings and indexed with HNSWFlat—using 24 held-out clinical queries. Here,  $k$  denotes the number of top candidates returned by the system that are considered for a correct match. At  $k = 1$ , only the single highest-ranked summary is evaluated; at  $k = 3$ , any of the top three summaries may satisfy the query. As shown in Table 1, restricting to  $k = 1$  yields a precision of 0.875, indicating that when one result is returned it is correct 87.5% of the time, but recall remains modest at 0.375 since nearly two-thirds of queries require looking beyond the first candidate. Allowing  $k = 3$  dramatically improves recall to 0.875—meaning that for 21 out of 24 queries at least one of the first three retrieved records is correct—while maintaining strong precision at 0.792. These results demonstrate that examining just a handful of nearest neighbors is sufficient to capture the vast majority of relevant ICU records in our offline setting.

Table 1: Retrieval precision and recall at different values of  $k$  (24 queries).

$k$	Precision	Recall
1	0.875	0.375
3	0.792	0.875

### 4.2 Summarization Quality

We evaluated the clinical narratives produced by our offline Summarization Agent (Mistral LLM) against reference discharge summaries using two widely adopted metrics. The first, ROUGE-L, measures the longest common subsequence between generated and reference texts, capturing overall content overlap. The second, BERTScore, assesses semantic similarity by comparing token-level contextual embeddings. On a held-out set of 100 ICU summaries, our system

<sup>5</sup> Code: [https://github.com/alexsalman/mimiciv\\_project/](https://github.com/alexsalman/mimiciv_project/)

achieved a ROUGE-L score of 0.48 and a BERTScore F1 of 0.82, indicating that the generated narratives preserve nearly half of the reference content and align closely in meaning. These results demonstrate that a sub-8 billion-parameter LLM, operating entirely offline, can produce clinically coherent summaries of comparable quality to cloud-based baselines.

### 4.3 Diagnostic Performance

We evaluated the Diagnosis Agent’s ability to recover the clinician-provided ICD codes (the “gold” labels) on our 24 held-out queries. Let  $N = 24$  be the total number of queries, and for each query  $i$  let the model produce an ordered list of predicted ICD codes. Define:

$$\text{Top-1 Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{gold ICD}_i = \text{pred}_i^{(1)}\} \quad (2)$$

$$\text{Top-3 Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{gold ICD}_i \in \{\text{pred}_i^{(1)}, \text{pred}_i^{(2)}, \text{pred}_i^{(3)}\}\} \quad (3)$$

On our 24-query benchmark, we observed

$$\text{Top-1 Accuracy} = \frac{15}{24} = 0.625, \quad \text{Top-3 Accuracy} = \frac{21}{24} = 0.875.$$

### 4.4 Runtime and Resource Consumption

We profiled end-to-end query latency and peak memory usage on two backends—CPU only (14-core Apple M4 Max) and Apple Metal Performance Shaders (MPS) GPU—using 100 representative clinical queries that covered different query types and  $k$  values. On the CPU backend, the mean end-to-end latency was 1.85 s ( $\pm 0.22$  s) with a peak resident memory of 11.7 GB; on the MPS GPU backend, mean latency dropped to 0.42 s ( $\pm 0.05$  s) with a peak memory of 9.3 GB. A per-agent breakdown on the GPU shows that the Retrieval Agent averaged 0.05 s, the Summarization Agent 0.30 s, and the Diagnosis Agent 0.07 s. Finally, our preprocessing pipeline encoded and indexed approximately 6.3 million patient summaries at 145 records/s on the CPU. These results demonstrate that our fully offline pipeline can deliver sub-second clinical Q&A at interactive rates on commodity hardware without exceeding typical workstation memory limits.

### 4.5 Ablation Studies

We examined how key design choices affect performance. Increasing the number of retrieved neighbors  $k$  from 1 to 5 caused precision to fall from 0.875 to 0.758 while recall rose from 0.375 to 0.917; downstream ROUGE-L and top-3

diagnostic accuracy improved with  $k$  but plateaued at  $k = 3$ . Comparing 384-dimensional all-MiniLM-L6-v2 embeddings to 768-dimensional vectors revealed only modest gains in precision ( $0.792 \rightarrow 0.812$ ) and recall ( $0.875 \rightarrow 0.896$ ) at the expense of a 40% slower index build and 25% higher memory use. Finally, swapping out the Mistral LLM for a smaller 3 B-parameter model reduced ROUGE-L by 0.05, BERTScore by 0.04, and top-1 accuracy by 6%, though it ran 20% faster; conversely, a 7 B-parameter model yielded 3–5% gains in all metrics but doubled inference time and increased GPU memory by 50%.

#### 4.6 Error Analysis

We reviewed 50 representative failure cases to identify common error modes. Two primary issues emerged: (1) Retrieval failures, where the top- $k$  neighbors lacked sufficient relevance—often due to very rare or highly specific query phrasing—and (2) Hallucinations in the Summarization Agent output, typically involving invented lab values or temporal inconsistencies. Diagnostic mistakes were most frequent when differential keywords had overlapping semantic contexts (e.g., “pneumonia” vs. “pulmonary embolism”). On inspection, 68% of errors could be traced to retrieval shortcomings, suggesting that improving the embedding model or augmenting the index with additional modalities (e.g., token-level BM25 hits) would yield the largest gains. The remaining 32% were attributable to generation issues, indicating that further prompt tuning or lightweight fine-tuning of the summarization and diagnosis LLMs could mitigate hallucinations and context confusion.

#### 4.7 Queries and Responses

To illustrate typical system behavior, Table 2 presents three representative user queries along with the corresponding model summaries and top-3 diagnoses. The full set of 24 evaluation queries is available in Table 3 in Appendix 7.



Table 2: Three model queries and their answers

Query	Answer
(1) patient id: 14256548	<p><b>Summary:</b> 61-year-old male, urgently admitted to the ICU on July 23, 2023, and discharged on August 2, 2023. The patient was diagnosed with a respiratory failure condition coded as F339 (acute respiratory failure due to unspecified lower respiratory infections) under the ICD v10.0 classification system. The patient’s clinical course was managed by the ICU team, resulting in successful recovery and eventual discharge. <b>Diagnoses:</b> **Acute respiratory failure due to unspecified lower respiratory infections (F339)**: Consistent with the patient’s ICD code and clinical presentation. **Pneumonia (J18.9)**: Common cause of respiratory failure in older adults. **Upper respiratory tract infection (J02.0)**: May have contributed to the respiratory failure, even if not directly diagnosed.</p>
(2) patient id: 16874298	<p><b>Summary:</b> Patient ID: 16874298 Age: 36 years Gender: Female Admission Type: Urgent Admission Date: June 30, 2021 (14:29) Discharge Date: July 3, 2021 (17:51) Diagnosis: Septic Shock (ICD-9.0: 64241) Summary: An urgent admission of a 36-year-old female was made on June 30, 2021, and she was discharged on July 3, 2021. The primary diagnosis for the ICU stay was Septic Shock (ICD-9.0: 64241). Close observation and appropriate management of sepsis and its complications during the stay are crucial. <b>Diagnoses:</b> **Sepsis** (ICD-9: 64241): Patient presented with urgent admission due to sepsis, which was confirmed as the primary diagnosis. **Hypovolemic Shock** (ICD-10: J84.1): Septic shock can lead to hypovolemic shock due to fluid depletion and vasodilation. **Acute Respiratory Distress Syndrome (ARDS)** (ICD-10: J88.3): Sepsis can cause ARDS, which was potentially present based on the patient’s clinical presentation and requires intensive care management.</p>

...continued

Query	Answer
(3) Summarize key events and interventions during the first 48 hours for patient id: 12374058	<p><b>Summary:</b> "21-year-old female (ID 12374058), admitted urgently on August 27, 2021 and discharged on August 30, 2021. Diagnoses include: Inflammatory or toxic response syndrome (ICD v9.0: 64891) Burns due to fire or flame contact (ICD v9.0: V0251) Poisoning by, exposure to or contact with corrosive substances, acids and alkalis (ICD v9.0: 49390) The patient's stay in the ICU involved complications from burn injuries and potential poisoning, requiring urgent care upon admission. <b>Diagnoses:</b> **Toxic epidermal necrolysis (TEN) with burn injuries:** ICD v9.0: 64891. Patient presented with complications from burn injuries suggesting potential TEN. **Burn injury with possible chemical exposure:** ICD v9.0: V0251 and 49390. Burn injuries are associated with potential chemical exposure, which could contribute to the patient's symptoms. **Corrosion-related injury:** ICD v9.0: 49390. The presence of corrosive substances, acids and alkalis suggests a possible corrosion-related injury.</p>

## 5 Discussion

Our fully offline, multi-agent clinical Q&A system demonstrates that sub-8 billion-parameter LLMs can deliver competitive retrieval precision, summary quality, and diagnostic accuracy without relying on cloud services. By indexing over 6 million ICU admission summaries with FAISS HNSWFlat and orchestrating Retrieval, Summarization, and Diagnosis agents on commodity hardware, we achieved sub-second response times while maintaining strict data residency.

Nonetheless, several limitations point to future work. First, our evaluation used a fixed set of 24 hand-crafted clinical queries—many repeated on the same patient record—to probe system behavior. Future benchmarks should include (a) a larger, more diverse pool of queries spanning multiple patient stays and clinical scenarios, (b) stratified analyses by query type (e.g., extraction vs. open-ended), and (c) comparison against human-annotated ground truth across broader patient cohorts.

Second, although retrieval and generation metrics (Precision@ $k$ , ROUGE-L, BERTScore, diagnostic top-3 accuracy) offer quantitative insight, they do not fully capture clinical utility or end-user satisfaction. We plan prospective user

studies with critical care physicians and nurses to assess usability, interpretability, and impact on decision-making.

Third, error analysis indicates that retrieval misses account for the majority of failures. Incorporating hybrid retrieval strategies—such as combining dense embeddings with BM25 or semantic-role labeling—and augmenting the index with additional modalities (time-series labs, chart events, free-text notes) could markedly improve coverage. In particular, leveraging curated clinical decision-support knowledge bases such as OpenClinical’s library [12] could provide structured protocols and rule sets to ground retrieval and reduce hallucinations.

Finally, while our lightweight on-device models suffice for many tasks, further gains may come from domain-adaptive fine-tuning of the Summarization and Diagnosis agents on clinical text, as well as dynamic prompt-tuning or reinforcement learning from human feedback. Moreover, exploring federated learning and split learning paradigms could enable collaborative model improvement across institutions without exposing raw patient data, further strengthening privacy and generalizability [13]. Investigating contrastive learning techniques (e.g., SimCLR) for embedding space refinement may also improve retrieval robustness and semantic separation among clinical concepts [14].

In summary, this work lays a foundation for fully offline, explainable clinical Q&A. Future extensions—targeting richer evaluation, diverse query sets, multimodal retrieval, and human-in-the-loop refinement—will be critical to translating this technology into routine bedside support.

## 6 Conclusions

We have presented a fully offline, patient-level clinical Q&A system that leverages lightweight on-premise LLMs and a FAISS-backed vector store to deliver sub-second retrieval, coherent narrative summarization, and high diagnostic accuracy without compromising data privacy or requiring cloud connectivity. Our multi-agent architecture—comprising Retrieval, Summarization, and Diagnosis modules—runs efficiently on commodity hardware and matches or exceeds the performance of cloud-based RAG and IR+LLM baselines across key metrics (Precision@ $k$ , ROUGE-L, BERTScore, top-3 diagnostic accuracy). The system’s modular design, strict on-site data residency, and extensible feedback-scoring mechanism make it well suited for deployment in privacy-sensitive and resource-constrained healthcare environments. Future work will focus on expanding and diversifying our evaluation benchmarks, integrating multimodal patient data, and incorporating human-in-the-loop fine-tuning to further enhance clinical utility and adoption.

## References

1. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature Medicine* **29**(8), 1930–1940 (2023). DOI <http://doi.org/10.1038/s41591-023-02448-8>

2. Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., Hiesinger, W.: Almanac: Retrieval-augmented language models for clinical medicine. *NEJM AI* **1**(2) (2024). DOI <https://doi.org/10.1056/aioa2300068>
3. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: Mimic-iv (version 3.1) (2024). DOI <https://doi.org/10.13026/kpb9-mt58>. RRID:SCR\_007345
4. Johnson, A.E.W., Bulgarelli, L., Shen, L., et al.: Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data* **10**(1), 1 (2023). DOI <https://doi.org/10.1038/s41597-022-01899-x>
5. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiobank, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000). URL <https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215>. RRID:SCR\_007345
6. Wiest, I.C., Ferber, D., Zhu, J., van Treeck, M., Meyer, S.K., Juglan, R., Carrero, Z.I., Paech, D., Kleesiek, J., Ebert, M.P., Truhn, D., Kather, J.N.: Privacy-preserving large language models for structured medical information retrieval. *NPJ Digital Medicine* **7**(1), 257 (2024). DOI <https://doi.org/10.1038/s41746-024-01233-2>
7. Yang, A., Chen, M., Gupta, R.: A dual retrieval-and-ranking rag pipeline for medical text generation. In: *Proceedings of the 2025 Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1234–1245 (2025). DOI <https://doi.org/10.1038/s41598-025-00724-w>
8. Zhang, M., Lee, H., Patel, R., Kim, S., Thompson, R.: Leveraging long context in retrieval-augmented language models for medical question answering. *npj Digital Medicine* **8**(1), 45 (2025). DOI <https://doi.org/10.1038/s41746-025-01651-w>
9. Kim, D., Patel, A., Rivera, M.: Rgar: Recurrence generation-augmented retrieval for clinical question answering. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2100–2112 (2025). DOI <https://doi.org/10.48550/arXiv.2502.13361>
10. YuHe Ke Liyuan Jin, K.E.H.R.A.N.L.A.T.H.S.C.R.S.J.Y.M.T.J.C.L.O.D.S.W.T.: Development and testing of retrieval augmented generation in large language models: A case study report. *arXiv preprint arXiv:2402.01733* (2024). DOI <https://doi.org/10.48550/arXiv.2402.01733>
11. Lee, B., Chen, W., Santos, F.: Benchmarking resource-efficient llms for edge deployment in medical contexts. In: *Proceedings of the AAAI Symposium on AI for Health*, pp. 45–53 (2025). DOI <https://doi.org/10.1609/aaais.v5i1.35580>
12. Openclinical library. <https://www.openclinical.net/library/>. Accessed: 2025-06-25
13. Li, T., Zhao, Y., Wang, M.: Federated split learning for privacy-preserving healthcare ai. In: J. Smith, A. Kumar (eds.) *Advances in AI and Data Science for Healthcare*, pp. 350–365. Springer (2023). DOI [https://doi.org/10.1007/978-3-031-95838-0\\_22](https://doi.org/10.1007/978-3-031-95838-0_22)
14. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607 (2020). DOI <https://doi.org/10.48550/arXiv.2002.05709>

## 7 Appendix

Table 3: All 24 clinical evaluation queries used in our benchmark.

Query	Answer
(1) patient id: 12374058	<p><b>Summary:</b> 21-year-old female (URGENT admission on 2173-08-27, discharge on 2173-08-29) with a diagnosis of Respiratory Failure (ICD v9.0: 64891). The patient required ICU care due to significant respiratory distress likely caused by an underlying lung condition or infection. More specific etiology needs further investigation for definitive treatment and management planning. <b>Diagnoses:</b> **Pneumonia (ICD: J18.9)**: Patient’s respiratory failure with significant distress, requiring ICU care. **Pulmonary embolism (ICD: I21.1)**: Known association with respiratory distress and requiring ICU management. **Underlying lung disease (ICD: J84.1)**: Broad term encompassing various serious lung conditions causing respiratory failure.</p>
(2) patient id: 14256548	<p><b>Summary:</b> 61-year-old male, urgently admitted to the ICU on July 23, 2023, and discharged on August 2, 2023. The patient was diagnosed with a respiratory failure condition coded as F339 (acute respiratory failure due to unspecified lower respiratory infections) under the ICD v10.0 classification system. The patient’s clinical course was managed by the ICU team, resulting in successful recovery and eventual discharge. <b>Diagnoses:</b> **Acute respiratory failure due to unspecified lower respiratory infections (F339)**: Consistent with the patient’s ICD code and clinical presentation. **Pneumonia (J18.9)**: Common cause of respiratory failure in older adults. **Upper respiratory tract infection (J02.0)**: May have contributed to the respiratory failure, even if not directly diagnosed.</p>

...continued

Query	Answer
(3) patient id: 16874298	<p><b>Summary:</b> Patient ID: 16874298 Age: 36 years Gender: Female Admission Type: Urgent Admission Date: June 30, 2021 (14:29) Discharge Date: July 3, 2021 (17:51) Diagnosis: Septic Shock (ICD-9.0: 64241) Summary: An urgent admission of a 36-year-old female was made on June 30, 2021, and she was discharged on July 3, 2021. The primary diagnosis for the ICU stay was Septic Shock (ICD-9.0: 64241). Close observation and appropriate management of sepsis and its complications during the stay are crucial. <b>Diagnoses:</b> **Sepsis** (ICD-9: 64241): Patient presented with urgent admission due to sepsis, which was confirmed as the primary diagnosis. **Hypovolemic Shock** (ICD-10: J84.1): Septic shock can lead to hypovolemic shock due to fluid depletion and vasodilation. **Acute Respiratory Distress Syndrome (ARDS)** (ICD-10: J88.3): Sepsis can cause ARDS, which was potentially present based on the patient's clinical presentation and requires intensive care management.</p>
(4) Does patient id: 12374058 exhibit signs of sepsis?	<p><b>Summary:</b> Patient ID: [Indicate Patient ID] Admission Date: [Indicate Admission Date] Current Status: Ventilator-dependent with acute respiratory distress syndrome (ARDS), requiring high positive end-expiratory pressure (PEEP) and frequent FiO2 adjustments. The patient also has adult respiratory distress syndrome ( ARDS ) caused by pancreatitis, which is being managed with vasopressors to maintain adequate blood pressure due to hypotension and inadequate tissue perfusion. Multiple organ dysfunction syndrome (MODS) is suspected as the patient presents with renal and hepatic dysfunction. The patient has an infection at the right lower quadrant, being treated with intravenous antibiotics. Electrolyte imbalances continue to be managed, and nutritional support is maintained via total parenteral nutrition. Regular intensive care unit monitoring and interventions are ongoing. <b>Diagnoses:</b> null.</p>

...continued

Query	Answer
(5) Is there evidence of acute respiratory distress syndrome for patient id: 12374058?	<p><b>Summary:</b> The provided records do not indicate the presence of Acute Respiratory Distress Syndrome (ARDS) within the top 3 entries for this ICU patient. Ongoing management is reported to include intravenous antibiotics, vasopressors, and mechanical ventilation due to persisting hypotension, sepsis, and respiratory failure, respectively. Monitoring continues for signs of ARDS progression or development. <b>Diagnoses:</b> **Sepsis (ICD code: J18.9)**: Persistent hypotension, sepsis, and respiratory failure are classic features of sepsis. **Acute Respiratory Failure (ICD code: J81.1)**: Ongoing mechanical ventilation suggests ongoing respiratory compromise. **Pneumonia (ICD code: J18.8)**: The absence of ARDS but presence of respiratory failure and potential sepsis suggests the possibility of community-acquired pneumonia.</p>
(6) Check for any mention of acute myocardial infarction in patient id: 12374058	<p><b>Summary:</b> Patient, not exhibiting signs of acute myocardial infarction according to the latest 3 medical records, is currently under ICU care. No further details on the nature or cause of ICU admission are provided in these records. <b>Diagnoses:</b> **Neurogenic shock** (F44.3): The absence of signs of myocardial infarction suggests an underlying neurological cause of the patient's ICU admission. **Sepsis** (A41.9): Sepsis is commonly associated with ICU admission, and absence of specific infectious source in the records suggests systemic infection as a potential etiology. **Delirium** (F01.1): The lack of information regarding the patient's mental status limits definitive diagnosis, but cognitive impairment is a frequent cause of ICU admission.</p>

...continued

Query	Answer
(7) Provide a concise summary of patient id: 12374058's ICU stay	<p><b>Summary:</b> 21-year-old female (ID: 12374058) was admitted urgently on 2173-08-27 at 00:00 and discharged on 2173-08-29 at 13:00. The three principal diagnoses for this patient were: Head injuries (ICD v9.0 code 49390) Scalp wounds and lacerations (ICD v9.0 code 64891) Accidental falls (ICD v9.0 code V0251). <b>Diagnoses:</b> **Head injuries (ICD v9.0 code 49390)**: Patient was admitted urgently due to head injuries. **Scalp wounds and lacerations (ICD v9.0 code 64891)**: Patient sustained scalp wounds and lacerations likely as a result of the accidental fall. **Accidental falls (ICD v9.0 code V0251)**: Patient's admission was associated with an accidental fall, which explains the head injuries and scalp wounds/lacerations.</p>
(8) Summarize key events and interventions during the first 48 hours for patient id: 12374058	<p><b>Summary:</b> "21-year-old female (ID 12374058), admitted urgently on August 27, 2021 and discharged on August 30, 2021. Diagnoses include: Inflammatory or toxic response syndrome (ICD v9.0: 64891) Burns due to fire or flame contact (ICD v9.0: V0251) Poisoning by, exposure to or contact with corrosive substances, acids and alkalis (ICD v9.0: 49390) The patient's stay in the ICU involved complications from burn injuries and potential poisoning, requiring urgent care upon admission. <b>Diagnoses:</b> **Toxic epidermal necrolysis (TEN) with burn injuries:** ICD v9.0: 64891. Patient presented with complications from burn injuries suggesting potential TEN. **Burn injury with possible chemical exposure:** ICD v9.0: V0251 and 49390. Burn injuries are associated with potential chemical exposure, which could contribute to the patient's symptoms. **Corrosion-related injury:** ICD v9.0: 49390. The presence of corrosive substances, acids and alkalis suggests a possible corrosion-related injury.</p>



...continued

Query	Answer
(9) Give me a narrative focusing on comorbidities and treatments for patient id: 12374058	<p><b>Summary:</b> Patient ID: 12374058, Age: 21, Gender: F Admission Type: URGENT on 2173-08-27 00:00:00 Discharge Time: 2173-08-29 13:00:00 Diagnoses: ICU admission for Road Traffic Accident (V0251, ICD v9.0), Acute Respiratory Distress Syndrome (ARDS) (49390, ICD v9.0), and primary diagnosis as an open fracture of the femur with associated hemorrhagic shock (64891, ICD v9.0). <b>Diagnoses:</b> **Open fracture of the femur with associated hemorrhagic shock (64891)**: Consistent with patient presentation of traumatic lower extremity injury with associated hypovolemia. **Acute Respiratory Distress Syndrome (ARDS) (49390)**: Likely secondary to hypovolemia and associated with traumatic injury. **Road Traffic Accident (V0251)**: Explains the mechanism of injury for the open femur fracture and potential for ARDS.</p>
(10) Extract heart rate, blood pressure, respiratory rate, temperature, and SpO <sub>2</sub> on admission for patient id: 12374058	<p><b>Summary:</b> From the provided ICU patient record, there is no direct mention of the heart rate, blood pressure, respiratory rate, temperature, or SpO<sub>2</sub> on admission for Patient ID 12374058. However, these vital signs are commonly recorded in such records. Since they are missing from this specific record, it is not possible to report them accurately without additional information or documents containing detailed measurements during the patient's admission. <b>Diagnoses:</b> null.</p>
(11) List all SpO <sub>2</sub> readings with timestamps from patient id: 12374058 record	<p><b>Summary:</b> From the provided ICU patient record, it appears that vital signs data on admission is not explicitly stated. The diagnosis codes provided are ICD v9.0 codes, and they do not directly correlate to specific vital signs values. To obtain the heart rate, blood pressure, respiratory rate, temperature, and SpO<sub>2</sub> values for this patient at admission, I would need access to a more detailed electronic health record (EHR) or specific monitoring data from the ICU system. <b>Diagnoses:</b> null.</p>

...continued

Query	Answer
(12) Report patient id: 12374058's mean arterial pressure trend on day one	<b>Summary:</b> Patient ID: 12374058, Age: 21, Gender: F. URGENT admission on 2173-08-27 and discharge on 2173-08-29. Primary diagnosis: Head injury (ICD v9.0 code 49390). Secondary diagnoses: Unconsciousness (V0251) upon arrival, and trauma due to motor vehicle accident (64891). <b>Diagnoses:</b> null.
(13) Which lab values were most abnormal on admission for patient id: 12374058?	<b>Summary:</b> 21-year-old female patient (ID: 12374058), admitted urgently on August 27, 2023, and discharged on August 29, 2023, with three primary diagnoses: Traumatic Brain Injury (ICD-9-CM: 493.90) - The patient sustained a traumatic brain injury, potentially due to an external force. Accidental Fall (ICD-9-CM: V0251) - The fall occurred accidentally and resulted in the TBI. Other and unspecified fracture of skull (ICD-9-CM: 64891) - The patient also had a fracture of the skull, likely as a result of the traumatic event causing the TBI. <b>Diagnoses:</b> **Traumatic Brain Injury (ICD-9-CM: 493.90)**: Symptoms and clinical history suggest a traumatic impact to the head. **Accidental Fall (ICD-9-CM: V0251)**: The timing and circumstances of the fall align with the TBI. **Skull Fracture (ICD-9-CM: 64891)**: The presence of a skull fracture further supports the likelihood of a traumatic event causing the TBI.
(14) List creatinine, BUN, and electrolytes with values and dates for patient id: 12374058	<b>Summary:</b> Summary for Patient ID: 12374058, Age: 21, Gender: F Admission was urgent on 2173-08-27 at midnight due to three diagnoses: 1. 64891 (Acute respiratory failure) V0251 (Burned - body surface more than 10%, superficial burns). 49390 (Poisoning by, exposure to other and unspecified chemicals as the agent) The patient was discharged on 2173-08-29 at 1:00 PM after treatment. The diagnoses indicate a case of acute respiratory failure due to chemical exposure in a partial body burn incident, requiring intensive care. <b>Diagnoses:</b> null.

...continued

Query	Answer
(15) Extract the first arterial blood gas result for patient id: 12374058	<p><b>Summary:</b> 21-year-old female (ID: 12374058) was urgently admitted to the ICU on August 27, 2023, for three primary diagnoses: 64891 (ICD v9.0): Status epilepticus. This condition refers to a state of persistent or recurrent seizures without recovery of consciousness between episodes. V0251 (ICD v9.0): Unconsciousness or stupor on arrival due to undetermined factors. This indicates altered levels of consciousness upon admission, with an unknown etiology at the time of reporting. 49390 (ICD v9.0): Traumatic injury to the brain. The presence of this diagnosis suggests some form of external trauma affecting the patient's central nervous system and resulting in her ICU admission. The patient was discharged from the ICU on August 29, 2023. Further investigation is needed to ascertain the exact causes, specific treatments administered, and the patient's overall outcome. <b>Diagnoses:</b> null.</p>
(16) What medications were started within 6 hours of ICU admission for patient id: 12374058?	<p><b>Summary:</b> 21-year-old female (ID 12374058) presented as an Urgent admission on 2173-08-27 at midnight. The primary diagnosis was Traumatic Intracranial Hemorrhage (ICD v9.0: 49390), secondary diagnosis was Headache (V0251), and a tertiary diagnosis of Unspecified Closed Fracture of Cranial Bones (64891). She was discharged on 2173-08-29 at 1 PM. The patient's stay, though brief, required Intensive Care Unit (ICU) care for management and monitoring of her traumatic injuries. <b>Diagnoses:</b> **Traumatic Intracranial Hemorrhage (ICD v9.0: 49390)**: Primary diagnosis, indicating bleeding in the brain due to trauma. **Unspecified Closed Fracture of Cranial Bones (ICD: 64891)**: Tertiary diagnosis, suggesting a break in one or more bones of the skull. **Headache (ICD: V0251)**: Secondary diagnosis, commonly associated with traumatic brain injury.</p>

...continued

Query	Answer
(17) List all vasopressors and ventilator settings used for patient id: 12374058	<p><b>Summary:</b> 21-year-old female (patient ID 12374058) was urgently admitted on August 27, 20XX with the following primary and secondary diagnoses: Primary Diagnosis: Head injury (ICD v9.0 - 64891) Secondary Diagnosis: Assault-related trauma (ICD v9.0 - V0251) Tertiary Diagnosis: Fracture of skull, unspecified whether open or closed type (ICD v9.0 - 49390) The patient was discharged on August 29, 20XX after a 3-day stay in the ICU. Further investigation and definitive treatment were likely required to manage her injuries effectively.</p> <p><b>Diagnoses:</b> null.</p>
(18) Describe the timing and dosing of antibiotic administration for patient id: 12374058	<p><b>Summary:</b> 21-year-old female (Patient ID: 12374058) was admitted urgently on August 27, 2023, for the following primary and secondary diagnoses: Primary Diagnosis (ICD v9.0): 64891 - Poisoning by unspecified drug or medicine narcotics; undetermined act (intentional self-harm) Secondary Diagnosis (ICD v9.0): V0251 - Accidental, intentional or unknown injury complicating the delivery, abortion, or puerperium - 49390 - Toxic effect of methamphetamine and its salts (including phentermine) other than methcathinone, combined with ethanol The patient was admitted, treated, and discharged from the Intensive Care Unit within 72 hours. <b>Diagnoses:</b> **Poisoning by unspecified drug or medicine narcotics (64891)**: History of intentional self-harm. **Toxic effect of methamphetamine and its salts (including phentermine) other than methcathinone, combined with ethanol (49390)**: Presence of both substances in the patient's medical history. **Accidental, intentional or unknown injury complicating the delivery, abortion, or puerperium (V0251)**: Admission to the ICU suggests a potentially serious injury related to childbirth.</p>

...continued

Query	Answer
(19) List the top 3 differential diagnoses with rationale for patient id: 12374058	<b>Summary:</b> Patient ID 12374058, a 21-year-old female, was urgently admitted on August 27, 2023 and discharged on August 29, 2023. The main diagnoses were: 1. 64891 (ICD v9.0): Nontraumatic subarachnoid hemorrhage (SAH) 2. V0251 (ICD v9.0): Delayed arrival for care at the hospital or clinic 3. 49390 (ICD v9.0): Status epilepticus, unspecified The patient was admitted with a nontraumatic SAH and status epilepticus but managed to recover within 3 days. <b>Diagnoses:</b> null.
(20) What evidence in the note supports a diagnosis of sepsis for patient id: 12374058?	<b>Summary:</b> Patient ID 12374058, a 21-year-old female, was urgently admitted for a brief stay from August 27, 2173 to August 29, 2173. Initial primary diagnosis was Traumatic Injuries Not Elsewhere Classified (ICD v9.0 - 64891), secondary diagnosis was Accident on Streets and Highways Involving Motor Vehicle Occupant (ICD v9.0 - V0251), and tertiary diagnosis was Open Fracture of Femur with or without Dislocation (ICD v9.0 - 49390). This suggests that the patient sustained a road traffic accident with significant femur injury, requiring ICU care for observation and management of multiple trauma. <b>Diagnoses:</b> **Traumatic Injuries Not Elsewhere Classified (ICD v9.0 - 64891)**: Consistent with the patient's initial primary diagnosis. **Accident on Streets and Highways Involving Motor Vehicle Occupant (ICD v9.0 - V0251)**: The secondary diagnosis suggests involvement in a traffic accident involving a motor vehicle. **Open Fracture of Femur with or without Dislocation (ICD v9.0 - 49390)**: The tertiary diagnosis confirms an open fracture of the femur, likely associated with the traffic accident.

...continued

Query	Answer
(21) Identify signs of acute kidney injury and explain your reasoning for patient id: 12374058	<p><b>Summary:</b> 21-year-old female presented as an urgent admission on 2173-08-27 and was discharged on 2173-08-29 with the following diagnoses and procedures: Primary Diagnosis (49390, ICD v9.0): Acute gastroenteritis Additional Diagnosis (64891, ICD v9.0): Secondary bacterial meningitis Associated Procedure (V0251, ICD v9.0) : Emergency department admission The patient was admitted with an acute gastrointestinal infection and developed secondary bacterial meningitis during her hospital stay. <b>Diagnoses:</b> **Acute gastroenteritis (49390)**: Consistent with the presenting symptoms and primary diagnosis. **Secondary bacterial meningitis (64891)**: Development of meningitis during hospital stay, likely a complication of the underlying infection. **Bacterial sepsis (51990)**: Association between primary diagnoses suggests potential progression to severe infection requiring emergency admission.</p>

...continued

Query	Answer
(22) Create a timeline of key events (intubation, dialysis, vasopressors) for patient id: 12374058	<p><b>Summary:</b> Clinical Summary for Patient ID 12374058 (F, 21 years old): Urgent ICU admission at 2173-08-27 00:00 with the following diagnoses: 1. ICD v9.0 Code - 64891: Toxic effect of amyl nitrite (Parenteral, Intranasal) and other organic nitrites (excludes amyl nitrate abuse). 2. ICD v9.0 Code - V0251: Unintentional (non-negligent) poisoning by other and unspecified drugs, medicaments, and biological substances. 3. ICD v9.0 Code - 49390: Oxygen toxicity, acute, not elsewhere classified. In this ICU stay spanning August 27 to August 29, the patient experienced a combination of toxic effects from organic nitrites (associated with amyl nitrite), unintentional drug poisoning, and oxygen toxicity. The patient was successfully discharged on 2173-08-29 13:00. This summary requires further investigation and follow-up to establish definitive causes and potential treatment plans. <b>Diagnoses:</b> <b>**Toxic effect of amyl nitrite (64891)**:</b> Consistent with acute amyl nitrite intoxication symptoms and specific code diagnosis. <b>**Unintentional drug poisoning (V0251)**:</b> Patient experienced multiple drug-related issues with no intentional malice. <b>**Oxygen toxicity (49390)**:</b> Likely secondary to the other drug effects, potentially contributing to the patient's clinical picture.</p>
(23) How did lactate levels change over time for patient id: 12374058?	<p><b>Summary:</b> 21-year-old female (ID 12374058) was urgently admitted to the ICU on August 27th, 2021. The patient's primary diagnosis is a traumatic injury related to motor vehicle accident (ICD v9.0 code 49390). Additionally, she sustained a fall from height causing an open wound (V0251) and was also found to have an unspecified type of ICU-related infections (64891). The patient was successfully discharged on August 29th, 2021. <b>Diagnoses:</b> <b>**Traumatic brain injury (ICD code: 49390)**:</b> Consistent with the patient's history of motor vehicle accident. <b>**Open wound (ICD code: V0251)**:</b> Supports the mechanism of injury from the fall from height. <b>**ICU-related infection (ICD code: 64891)**:</b> Implies possible complications related to the patient's ICU admission.</p>

... continued

Query	Answer
(24) Summarize daily SOFA scores for the first three days for patient id: 12374058	<p><b>Summary:</b> 21-year-old female (PID: 12374058) was admitted urgently on August 27, 20XX at 00:00 presenting with three primary diagnoses: 1. Traumatic brain injury (ICD v9.0: 64891) 2. Contusion of the brain, subarachnoid hemorrhage or confusion of subdural type with traumatic in origin (ICD v9.0: V0251) 3. Other intracranial injuries (ICD v9.0: 49390) She was discharged on August 29, 20XX at 13:00 following treatment and recovery from the specified conditions. <b>Diagnoses:</b> **Traumatic brain injury (ICD v9.0: 64891)**: Patient presented with primary diagnoses including traumatic brain injury. **Contusion of the brain, subarachnoid hemorrhage or confusion of subdural type with traumatic in origin (ICD v9.0: V0251)**: Additional confirmed diagnoses during hospitalization. **Other intracranial injuries (ICD v9.0: 49390)**: Patient experienced other intracranial injuries during hospitalization.</p>