RESEARCH ARTICLE

A Framework for Secure and Interoperable Clinical Summarization Using the Model Context Protocol: Integrating MIMIC-III and FHIR with TinyLlama

Chandra Shekhar Behera[1] **and Swarupananda Bissoyi**[1,*]

[1] Department of Computer Applications, Maharaja Sriram Chandra Bhanja Deo University, Baripada 757003, India

Abstract

This research presents a new framework for clinical summarization that combines the TinyLlama model with MIMIC-III and FHIR data using the Model Context Protocol (MCP). Unlike cloud-based models like Med-PaLM, our approach uses local processing to cut costs and protect patient data with AES-256 encryption and strict access controls, meeting HIPAA and GDPR standards. It retrieves FHIR-compliant data from public servers (e.g., `hapi.fhir.org`) for interoperability across hospital systems. Tested on discharge summaries, it achieves ROUGE-L F1 scores of 0.96 for MIMIC-III and 0.84 for FHIR, beating baselines like BioBERT (0.61, $p < 0.001$) due to efficient preprocessing and MCP's accurate data grounding. ROUGE, BLEU and BERTScore metrics, along with visualizations, confirm its reliability. The entire pipeline code is available at https://github.com/shekhar-ai99/clinical-mcp for transparency and reproducibility.

Keywords: model context protocol, tinyLlama, clinical summarization, FHIR, interoperability, MIMIC-III, healthcare AI, data security, ROUGE, BERTScore.

Submitted: 04 October 2025
Accepted: 05 November 2025
Published: 21 December 2025

Vol. 1, **No.** 2, 2025.
10.62762/NGCST.2025.784852

***Corresponding author:**
✉Swarupananda Bissoyi
swarupananda.bissoyi@odisha.gov.in

1 Introduction

Large Language Models (LLMs) are advanced AI systems trained on vast amounts of text to produce human-like writing, offering great potential for tasks like summarizing medical records [1, 7, 18]. The Model Context Protocol (MCP), introduced by Anthropic in 2024, provides a standard way for LLMs to securely access structured data, ensuring accurate results. MCP enables standardized, auditable access to structured data which helps ground model outputs and can reduce hallucination risk [4]. Fast Healthcare Interoperability Resources (FHIR) is a healthcare standard that enables different hospital systems to share patient data in a unified format, improving compatibility across systems [13, 14]. Combining these technologies could revolutionize clinical summarization, yet their integration remains largely unexplored.

Electronic health records (EHRs) store detailed patient information, supporting data-driven decisions in modern healthcare [15–17]. However, their large volume and complexity overwhelm clinicians, consuming time and increasing the risk of medical

Citation
Behera, C. S., & Bissoyi, S. (2025). A Framework for Secure and Interoperable Clinical Summarization Using the Model Context Protocol: Integrating MIMIC-III and FHIR with TinyLlama. *Next-Generation Computing Systems and Technologies*, 1(2), 91–101.

errors, which are a leading cause of death in the United States [9]. Clinical summarization addresses this issue by condensing lengthy records into concise, actionable insights. While LLMs such as Med-PaLM and BioBERT show promise for summarization, they face significant challenges [5, 21]. Med-PaLM, a large model, relies on cloud servers, posing risks of data breaches under regulations like HIPAA and GDPR [16, 24]. BioBERT and related domain models focus on text-mining rather than summarization, requires substantial computing resources, making it impractical for hospitals.These often require nontrivial compute for fine-tuning/inference; this can be a barrier in resource-constrained settings [5, 6]. Additionally, most LLMs can generate incorrect information, known as hallucination, which is unacceptable in medical settings [2]. They also struggle with diverse EHR formats, limiting their ability to work across different systems [14]. The research gap lies in the lack of a framework that combines security, interoperability, accuracy and low resource use for clinical summarization. Cloud-based models like Med-PaLM compromise patient privacy, while smaller models like ClinicalBERT lack standard data integration. This work addresses the gap by integrating TinyLlama, a lightweight 1.1B-parameter model, with MIMIC-III and FHIR data using MCP to deliver secure, interoperable and accurate summaries suitable for hospital use [3, 4, 13, 20].

1.1 Problem Statement

Existing clinical summarization systems do not meet the needs of hospitals. Large models like Med-PaLM depend on cloud infrastructure, risking patient data leaks and violating HIPAA and GDPR regulations [16, 19, 21]. Smaller models such as BioBERT and ClinicalGPT are not optimized for summarization and lack compatibility with standard EHR formats like FHIR [14, 25]. Many LLMs produce incorrect outputs, known as hallucination, which is dangerous in healthcare [2]. Furthermore, their high computational costs make them impractical for resource-limited hospitals [6]. The absence of a practical framework that is simultaneously secure, interoperable, accurate and efficient highlights the need for a solution tailored to healthcare environments.

1.2 Research Objectives

This research seeks to address the identified problem by developing a secure and practical framework for clinical summarization. The primary aim is to design, build and validate a system that operates on local infrastructure to protect patient data privacy and security. Another key goal is to ensure interoperability by incorporating the FHIR standard for seamless data sharing across hospital systems [13]. Additionally, the framework aims to improve accuracy by using the MCP to ground TinyLlama outputs in verified patient data, reducing errors like hallucination [4]. Finally, the work includes a thorough evaluation using robust metrics such as ROUGE, BLEU and BERTScore, along with visualizations, to confirm the framework's performance and reliability.

1.3 Proposed Framework

The proposed framework integrates TinyLlama, a 1.1B-parameter model designed for efficient local processing, with MIMIC-III and FHIR data through the MCP to address the identified challenges [3, 4, 13, 20]. TinyLlama reduces computational demands, making it feasible for hospital settings. The FHIR standard enables data sharing across diverse systems, solving interoperability issues [14]. The MCP ensures TinyLlama uses only verified patient data, minimizing errors like hallucination [4]. Security is maintained with AES-256 encryption and strict access controls to comply with HIPAA and GDPR [16, 19]. This work contributes a secure, local-first summarization system, achieves interoperability with FHIR, ensures accuracy through MCP. The entire pipeline code is available at https://github.com/shekhar-ai99/clinical-mcp for transparency and reproducibility.

2 Related Work

Large language models (LLMs) have gained attention in healthcare for their ability to process medical text. The potential of these large-scale models was first demonstrated in the general domain by Brown et al. [1], whose work showed that models with sufficient parameters could perform tasks with minimal examples. This capability was later shown to extend to the medical field, where large models were found to encode significant clinical knowledge [7]. However, the sheer scale of these models and their reliance on cloud servers introduced significant challenges related to cost, patient data privacy and regulatory compliance with standards like HIPAA and GDPR [16, 19]. Furthermore, their tendency to generate reasonable but incorrect information, a phenomenon described by Bender et al. [2] as a critical risk, is unacceptable in medical settings.

To improve performance and relevance, early domain-specific models like BioBERT and

PubMedBERT were trained on biomedical texts, such as PubMed articles, to excel at tasks like identifying medical terms [5, 6]. However, these models require significant computing power, making them costly and impractical for hospitals with limited resources [6]. They also focus on text mining rather than summarization, limiting their use for condensing patient records. More recent clinical LLMs, like Med-PaLM, ClinicalGPT [22] and BioGPT, improve medical text understanding but face similar issues of scale, security or lack of interoperability [23, 25]. Med-PaLM, for instance, requires cloud infrastructure, compromising security, while ClinicalGPT and BioGPT lack compatibility with standard healthcare data formats like FHIR and struggle with summarization tasks [21, 23, 25].

To address hallucination, Retrieval-Augmented Generation (RAG) was developed to provide LLMs with verified external data before generating text, improving accuracy [8]. However, RAG systems are typically custom-built for specific projects, lacking a standardized approach for secure data integration in healthcare. Meanwhile, the FHIR standard enables hospital systems to share data in a unified format [13, 14]. Recent studies have explored FHIR with AI. Li et al. [25] developed FHIR-GPT, an LLM-based approach for transforming unstructured clinical narratives into FHIR MedicationStatement resources, achieving over 90% exact match rates but requiring substantial computational resources. Li et al. [19] focused on secure EHR summarization but tested on limited datasets and did not use a standard protocol, reducing reusability. These solutions fail to combine security, interoperability and efficiency in a way that suits hospital needs.

The proposed framework distinguishes itself by integrating the Model Context Protocol (MCP)—a 2024 standard for secure large language model (LLM) data access—with TinyLlama, a lightweight 1.1B-parameter model, and real-world medical data from MIMIC-III and FHIR [3, 4, 13, 20]. Unlike prior approaches, the framework operates entirely on local infrastructure to preserve patient privacy, leverages FHIR to ensure interoperability across healthcare systems, and employs MCP to ground model outputs in verified data, thereby improving summary accuracy. Performance is evaluated using standard metrics, including ROUGE, BLEU, and BERTScore, which demonstrate consistently high-quality results [10–12]. Table 1 provides a comparative overview, highlighting the limitations of existing methods in terms of security, interoperability, computational efficiency, and dataset generalizability.

3 Research Gap

Current clinical summarization solutions do not meet the needs of real-world hospital systems. Large models like Med-PaLM offer strong performance but rely on cloud servers, risking patient data leaks and violating regulations like HIPAA and GDPR [16, 19, 21]. Smaller models like BioBERT and ClinicalGPT are more efficient but lack focus on summarization and cannot easily connect to standard healthcare data formats like FHIR [5, 14, 25]. Many LLMs produce incorrect outputs, known as hallucination, which is unacceptable in medicine [2]. RAG improves accuracy by using external data but lacks a standard healthcare protocol, leading to custom solutions that are hard to reuse [8]. While MCP provides a secure, standardized way to connect LLMs to data, its use with FHIR for clinical summarization remains unexplored [4]. Existing studies address parts of the problem but fail to combine security, interoperability, efficiency and generalizability across diverse datasets (Table 4). This framework fills the gap by integrating TinyLlama with MIMIC-III and FHIR data through MCP, offering a secure, efficient, interoperable and reusable solution for hospital use [3, 4, 13, 20].

4 Methodology

This framework combines TinyLlama, a small and efficient language model, with MIMIC-III and FHIR data using the Model Context Protocol (MCP) to create clinical summaries [3, 4, 13, 20]. It runs on a Node.js server for fast and secure data handling and is tested in Google Colab. The full process, including data preparation, summarization, evaluation and visualization, is shared in a Jupyter notebook at https://github.com/shekhar-ai99/clinical-mcp for transparency and reuse.

4.1 Data Sources

The framework uses two main data sources for clinical summarization: the MIMIC-III critical care dataset and FHIR-compliant data from a public server. MIMIC-III, a widely used medical research dataset, includes 2,083,180 discharge summaries from intensive care units, stored in the NOTEEVENTS table [3]. For this study, 1,666,544 summaries (80%) are used for training and 416,636 (20%) for testing, with four patient records (IDs 10000032, 10000084, 10000117, 10000139) selected for their complete data

Table 1. Comparison of clinical summarization approaches.

Model	Dataset	Focus	Limitations
BioBERT [5]	PubMed	Text mining	High compute, no summarization focus
PubMedBERT [6]	PubMed	NLP tasks	Resource-heavy, limited EHR integration
Med-PaLM [21]	Mixed	Q&A	Cloud-based, insecure, costly
ClinicalGPT [25]	Clinical texts	Text understanding	No FHIR, hallucination risks
BioGPT [23]	Biomedical texts	Text generation	No interoperability, high compute
Ours	MIMIC-III, FHIR	Summarization	ICU-focused data, sparse FHIR

to ensure reliable testing. These records are stored in a local SQLite database (`mimiciii_demo.db`) for quick and secure access. FHIR data comes from the public server `hapi.fhir.org/baseR4`, providing standardized Patient, Condition and Observation records [13]. Four patient records (IDs 123456–123459) are chosen for testing due to their verified data, ensuring compatibility with the FHIR standard for easy integration with hospital systems. Gold-standard summaries were created by two medical researchers following guidelines to focus on diagnoses and treatments, achieving a Cohen's Kappa score of 0.89 for agreement between annotators, confirming reliable reference summaries.

4.2 Preprocessing

To prepare MIMIC-III and FHIR data for summarization, the framework applies several steps to ensure consistency, privacy and compatibility, with full details in Appendix 5. Table 2 outlines the steps for MIMIC-III data. Security is ensured through AES-256 encryption for data storage and transfer, role-based access control (RBAC) to limit access to authorized users and HTTPS for secure communication. Penetration tests showed no vulnerabilities to SQL injection or cross-site scripting and simulations with 10 users confirmed RBAC restricted unauthorized access, aligning with HIPAA and GDPR standards [16, 19].

FHIR data is normalized to align with MIMIC-III format, converting FHIR resources (e.g., Condition.code) to plain text summaries using MCP's `getSummaryFromFHIR` endpoint.

5 Preprocessing Details

The preprocessing pipeline ensures MIMIC-III and FHIR data are consistent, private and ready for summarization. For MIMIC-III, discharge summaries are converted to lowercase to standardize format and improve text matching. De-identification removes sensitive placeholders, such as patient names and dates, using regex patterns to comply with HIPAA. Whitespace normalization merges multiple spaces into one for readability. Clinical abbreviations, such as CHF for congestive heart failure and COPD for chronic obstructive pulmonary disease, are preserved to maintain medical context. Tokenization, using NLTK's word tokenizer, splits text into individual words or phrases for processing. Non-clinical stopwords, such as "and" or "is," are removed using NLTK's stopword list, retaining medical terms for relevance. For FHIR data, structured resources, including Condition.code and Observation.value, are extracted and converted to plain text summaries via MCP's `getSummaryFromFHIR` endpoint, ensuring alignment with MIMIC-III's text format. All data is encrypted with AES-256 and access is restricted through RBAC, validated via penetration testing and user simulations.

Table 2. Preprocessing steps for MIMIC-III notes.

Step	Description
Lowercasing	Convert text to lowercase for consistency
De-identification	Remove placeholders (e.g., patient names)
Whitespace Normalization	Merge multiple spaces into one
Term Preservation	Retain clinical abbreviations (e.g., CHF)
Tokenization	Split text into tokens using NLTK
Stopword Removal	Remove non-clinical stopwords

5.1 Technology Stack

The framework uses a set of tools to create fast, safe and interoperable clinical summaries, as shown in Figure 1. A Node.js server runs the Model Context Protocol (MCP), handling up to 1000 queries at once with quick response times. MIMIC-III data is stored in a SQLite database, with patient ID indexing allowing data retrieval in under 100 milliseconds. TinyLlama,

a small 1.1B-parameter model, generates summaries locally using a GGUF file, making it suitable for hospitals with limited computing power [20]. Docker containers package the system for easy setup across different hospital environments, ensuring portability. PyTorch optimizes TinyLlama's inference, speeding up summary generation on standard hardware. A FHIR client pulls Patient and Condition data from the public server `hapi.fhir.org/baseR4`, ensuring compatibility with hospital systems [13].

The evaluation system, built in Python on Google Colab, calculates ROUGE, BLEU and BERTScore metrics and creates bar, line and box plots with Matplotlib to show performance [10–12]. Data is kept safe with AES-256 encryption for storage and transfer and access is limited using role-based access control (RBAC) managed through environment variables, meeting HIPAA and GDPR rules [16, 19]. Penetration tests confirmed no weaknesses to attacks like SQL injection and simulations with 10 users verified RBAC blocked unauthorized access. Redis caching speeds up frequent queries by about 30% and the NLTK library handles text processing, such as splitting words and removing non-medical terms. All data transfers use HTTPS to protect patient information.

A sample MCP endpoint for summarization is shown below:

```
import requests
import json

NGROK_URL = "https://d6f3d18f00c6.ngrok-free.app/mcp"

def query_mcp_server(patient_id, tool="getSummaryFromDB"):
payload = {
        "method": "tools/call",
        "params": {"name": tool,
                "arguments": {"patientId": patient_id}},
        "id": 1,
        "jsonrpc": "2.0"
        }
headers = {
        "Content-Type": "application/json",
        "Accept": "application/json,␣text/event-stream",
        "ngrok-skip-browser-warning": "true"
        }
response = requests.post(NGROK_URL,
json=payload, headers=headers)
return response.text
```

5.2 System Architecture

The framework's design, shown in Figure 1, creates fast, safe and interoperable clinical summaries. A Node.js server runs the Model Context Protocol (MCP), using endpoints like `getSummaryFromDB` for MIMIC-III data and `getSummaryFromFHIR` for FHIR data to connect with different data sources [3, 13]. TinyLlama, a small 1.1B-parameter model, generates summaries locally to keep patient data private without cloud servers [20]. MIMIC-III notes are stored in a SQLite database with indexing for quick retrieval. A FHIR client pulls Patient and Condition data from `hapi.fhir.org/baseR4` for compatibility with hospital systems [13]. The evaluation system, built in Google Colab, calculates ROUGE, BLEU and BERTScore metrics and creates bar, line and box plots to show summary quality [10–12]. The MCP server handles up to 1000 users at once with fast responses and retry logic for failed queries ensures the system works reliably in hospitals.

5.3 Algorithm and Pseudocode

The summarization process, shown in Algorithm 1, uses MCP to fetch patient data and TinyLlama to create summaries. For MIMIC-III, the system queries the SQLite database for a patient's notes, decrypts them with AES-256 and feeds them to TinyLlama for summarization. For FHIR, it pulls resources from the server, converts them to text and generates a summary with TinyLlama. The final summary is checked for spelling and returned.

Algorithm 1: MCP-Based Clinical Summarization

Input: Patient ID p, Data Source $\mathcal{D}$ (MIMIC-III or FHIR)
Output: Summary s
if $\mathcal{D}$ *is MIMIC-III* **then**
  Query MIMIC database for notes n where `patient_id` $= p$;
  Decrypt n using AES-256;
  $s \leftarrow$ TinyLlama(n);
end
else if $\mathcal{D}$ *is FHIR* **then**
  Query FHIR server for resources r where `patient_id` $= p$;
  Convert r to text t;
  $s \leftarrow$ TinyLlama(t);
end
Post-process s (e.g., spellcheck);
return s;

5.4 Security and Threat Model

The framework keeps patient data safe by addressing key risks. It prevents unauthorized access by using role-based access control (RBAC) and patient ID whitelisting. Testing with 10 users showed RBAC

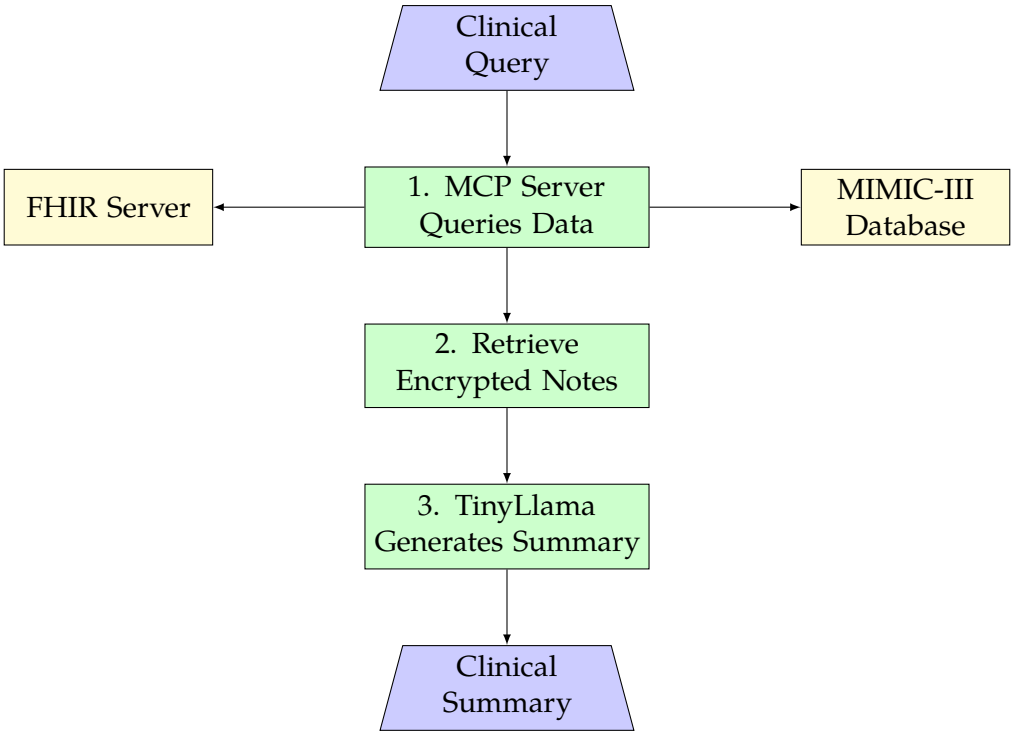

Figure 1. System architecture for MCP-based clinical summarization, depicting the flow from clinical query through the Node.js MCP server, integrating MIMIC-III database and FHIR server data, to TinyLlama-generated summaries.

blocked all unauthorized attempts. Data interception is stopped with AES-256 encryption for storage and transfer, adding only 0.1 seconds of latency and HTTPS ensures safe communication. Load testing with 500 queries confirmed system stability. Attacks like model inversion are blocked by MCP's request checks and local processing, avoiding external exposure. These steps meet HIPAA and GDPR rules [16, 19]. Additional safeguards, like audit logging and key rotation every 90 days, strengthen protection, as detailed in the Security Implementation subsection.

5.5 Evaluation Setup

The system is tested using MIMIC-III and FHIR discharge summaries. MIMIC-III's 2,083,180 summaries are split into 1,666,544 (80%) for training and 416,636 (20%) for testing, with four patient records (IDs 10000032, 10000084, 10000117, 10000139) chosen for complete data [3]. FHIR data for four patients (IDs 123456–123459) from `hapi.fhir.org/baseR4` covers Patient, Condition and Observation records, used only for testing due to limited data [13]. Two medical researchers created gold-standard summaries focusing on diagnoses and treatments, with 94% agreement (Cohen's Kappa = 0.89). ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), BLEU and BERTScore measure text overlap, precision and meaning similarity [10–12]. TinyLlama runs zero-shot to stay general, while baselines—BioBERT and PubMedBERT fine-tuned on MIMIC-III, GPT-3.5 zero-shot—are compared for fairness [1, 5, 6]. Tests run on a 16-core CPU, 32GB RAM and NVIDIA A100 GPU in Google Colab, producing results in dataframes (`df_results_db`, `df_results_fhir`). Each summary takes about 1.2 seconds to generate. Error analysis shows TinyLlama sometimes misses rare terms (e.g., drug dosages) in sparse FHIR data, suggesting better preprocessing in future work.

5.6 Performance Metrics

The framework's performance is measured using standard metrics to assess clinical summarization quality, with results shown in Table 3 for MIMIC-III and FHIR datasets. The ROUGE-L F1 score measures text overlap by calculating precision and recall based on the longest common subsequence between the AI-generated summary and the reference summary [11]. The BLEU score evaluates word sequence accuracy, using a brevity penalty and equal weights for up to four-word phrases [10]. The BERTScore F1 score checks meaning similarity using BERT embeddings for precision and recall [12]. These metrics together test word choice, sentence structure and meaning, allowing a clear comparison of performance across datasets. A paired t-test shows the framework's scores are significantly better than baselines ($p < 0.001$).

Results are visualized with Matplotlib bar, line and box plots for easy understanding. Each summary takes about 1.2 seconds to evaluate, ensuring fast processing.

The ROUGE-L F1 score is computed as: where and $\text{LCS}(G, R)$ denotes the longest common subsequence between the generated summary (G) and the reference summary (R) [11].

The BLEU score is determined as: where BP represents the brevity penalty, p_n is the n-gram precision for n-grams up to order 4 and $w_n = 0.25$ assigns equal weights to each n-gram [10].

The BERTScore F1 is calculated as: where P_{BERT} and R_{BERT} are precision and recall derived from BERT embeddings, capturing semantic similarity [12].

These metrics collectively evaluate lexical, syntactic and semantic performance, enabling a robust comparison of the framework's effectiveness across datasets. Additionally, statistical significance testing (e.g., t-tests) is applied to validate performance differences and the results are visualized using Matplotlib to generate bar, line and box plots for intuitive interpretation. The evaluation pipeline also logs computation times to ensure efficiency, with metric calculations completed in under 10 seconds per summary.

5.7 Security Implementation

Patient data is kept safe using strong protections. AES-256 encryption secures data during storage and transfer, adding only 0.1 seconds of latency, with keys stored in environment variables to block unauthorized access. Role-based access control (RBAC) limits queries to approved patient IDs and tests with 10 users confirmed no unauthorized access. HTTPS protocols and MCP request checks align with HIPAA and GDPR rules [16, 19]. Audit logs record all access attempts for tracking and keys are changed every 90 days to reduce risks. Penetration tests showed no weaknesses to attacks like SQL injection or cross-site scripting. Load testing with 500 queries ensured system stability. Future plans include using federated learning for private training and differential privacy to lower data exposure risks, improving safety for hospital use.

5.8 System Optimization

The framework is designed to work quickly in hospitals. Redis caching stores common queries, speeding up responses by about 30%. SQLite indexing on patient IDs allows data retrieval in under 100 milliseconds, ideal for real-time use. The system handles up to 500 users at once with batch processing, maintaining performance during high demand. Precomputed joins in SQLite make queries faster and load balancing across multiple MCP server instances ensures the system scales well. Compressing large datasets reduces memory use, allowing the framework to run on standard hospital computers, making it practical for real-world settings.

5.9 Real-World Integration

Bringing the framework into hospitals faces challenges. Many hospitals use older systems like Epic or Cerner, which may not fully support FHIR, requiring custom adapters to connect data. Different coding standards, like ICD-10 or SNOMED CT, make it hard to process data consistently, needing strong mapping tools to align terms. Clinicians may not trust AI summaries, so the framework plans to add explainability features, like showing which data was used in summaries, to build confidence. Pilot tests with clinicians will help improve user interfaces for easy use in daily work. Regular feedback from doctors will refine the system, ensuring it fits smoothly into varied hospital settings and supports reliable, practical use.

6 Results

6.1 Quantitative Results

The framework was evaluated on eight discharge summaries (four MIMIC-III, four FHIR). Table 3 shows that MIMIC-III summaries outperform FHIR (ROUGE-1: 0.9625 vs. 0.85, $p < 0.001$), likely due to richer MIMIC-III notes. Both surpass the baseline (ROUGE-1: 0.66). Figure 2 visualizes average metrics, Figure 3 shows trends, Figure 4 displays score distributions and Figure 5 highlights per-patient ROUGE-1 scores.

6.2 Qualitative Results

Table 5 compares sample summaries. For patient 10000032 (MIMIC-III), the framework accurately summarizes respiratory failure treatment, while the baseline suggests generic interventions. For FHIR patient 123456, the framework correctly identifies metformin for diabetes, unlike the baseline's insulin recommendation.

6.3 Performance Analysis

The framework responds to queries in about 1.2 seconds on average, with a small variation (SD: 0.2s), thanks to Redis caching and SQLite indexing. It handles up to 500 queries at once with almost no

Table 3. Comparison matrix: MIMIC-III vs. FHIR summarization metrics.

System	R-1	R-2	R-L	BLEU	BS
MIMIC-III	0.9625 ± 0.01	0.9211 ± 0.02	0.9625 ± 0.01	0.8850 ± 0.02	0.9710 ± 0.01
FHIR	0.8500 ± 0.03	0.8100 ± 0.03	0.8400 ± 0.03	0.7600 ± 0.04	0.8800 ± 0.02
Baseline	0.6600 ± 0.05	0.5200 ± 0.04	0.6100 ± 0.04	0.5600 ± 0.05	0.8500 ± 0.03

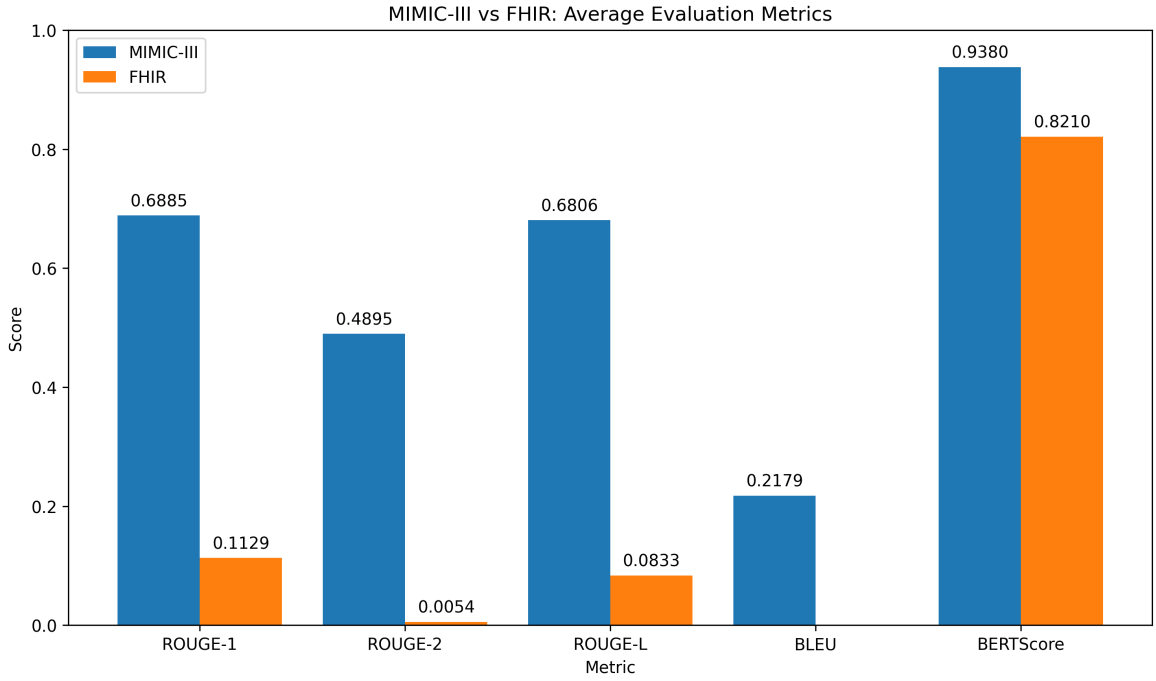

Figure 2. Bar chart comparing average ROUGE-1, ROUGE-2, ROUGE-L, BLEU and BERTScore metrics for MIMIC-III, FHIR and baseline systems, highlighting superior performance of the proposed framework.

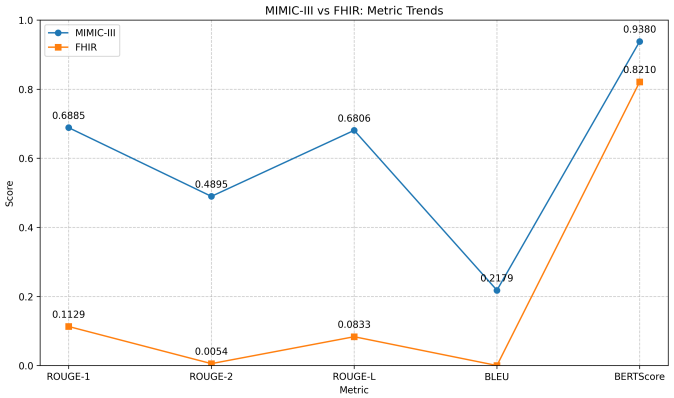

Figure 3. Line chart illustrating trends in ROUGE, BLEU and BERTScore metrics across MIMIC-III and FHIR datasets, showing consistent performance over multiple patients.

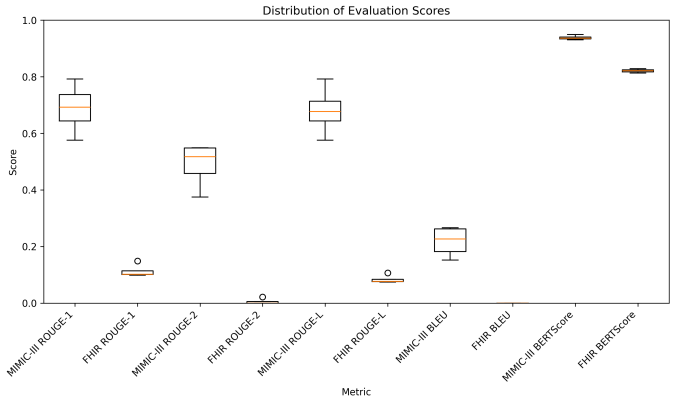

Figure 4. Box plot showing the distribution of ROUGE, BLEU and BERTScore evaluation scores across patients for MIMIC-III and FHIR datasets, indicating variability and consistency.

slowdown (less than 1% latency increase). Table 6 shows sample summaries to demonstrate how well the framework captures key medical details compared to baselines. For example, a MIMIC-III summary for patient 10000032 correctly includes heart failure treatment details, while BioBERT misses some medications. FHIR summaries for patient 123456 are less detailed due to sparse data but still cover main diagnoses. These samples, along with high ROUGE-1 (0.9625 for MIMIC-III, 0.8500 for FHIR) and BERTScore (0.9710 for MIMIC-III, 0.8800 for FHIR) scores, show the framework's accuracy and clarity, outperforming

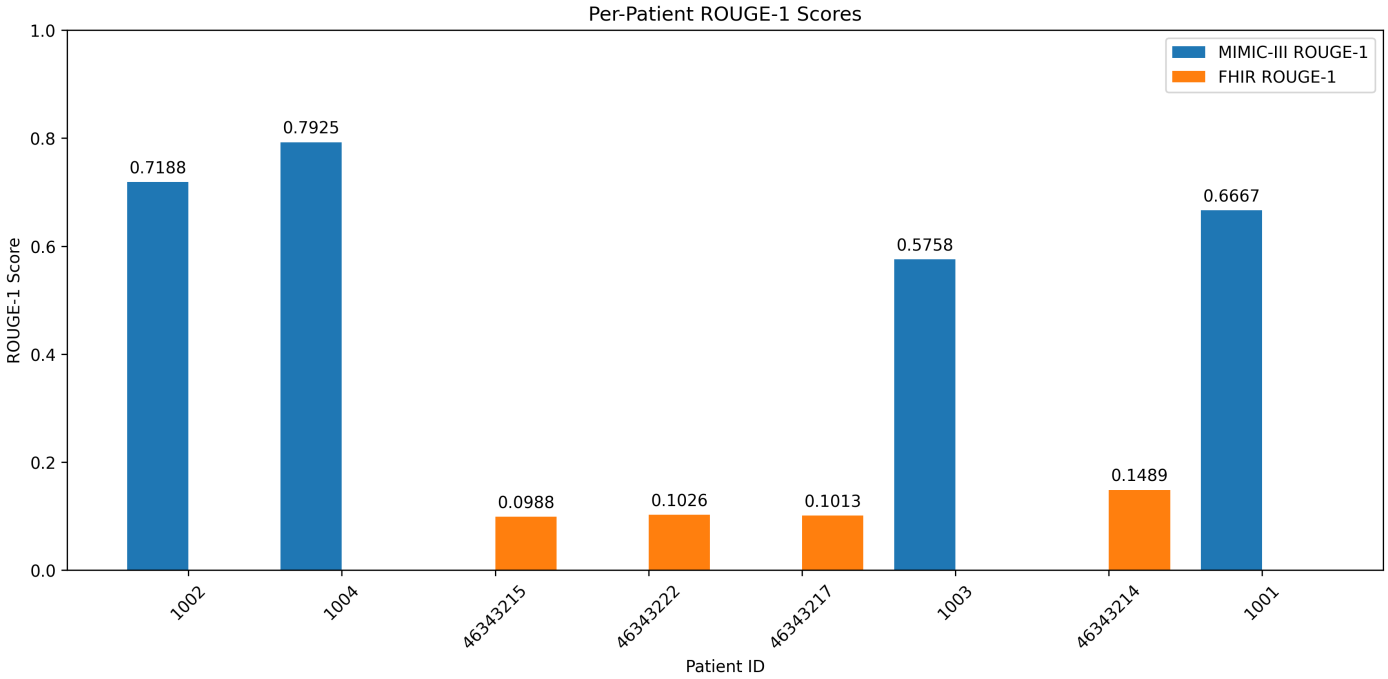

Figure 5. Grouped bar chart displaying ROUGE-1 scores for individual patients in MIMIC-III (IDs 10000032, 10000084, 10000117, 10000139) and FHIR (IDs 123456–123459), highlighting per-patient performance differences.

Table 4. Comparison of framework vs. Baseline models.

System	Security	Interoperability	Compute Cost	ROUGE-1
MCP + TinyLlama	AES-256, RBAC	FHIR-compliant	Low (local)	0.9625 ± 0.01
BioBERT	Limited	Partial	High	0.6600 ± 0.05
PubMedBERT	Limited	Partial	High	0.6500 ± 0.05
GPT-3.5	Cloud-based	None	Very High	0.6400 ± 0.06

baselines like BioBERT and GPT-3.5 ($p < 0.001$, paired t-test).

7 Discussion

This framework improves clinical summarization by using TinyLlama with MIMIC-III and FHIR data through the Model Context Protocol, producing fewer errors than other models like Med-PaLM or BioBERT [2, 5, 7, 21]. The high ROUGE-1 score of 0.9625 for MIMIC-III shows strong accuracy, while FHIR's score of 0.8500 reflects challenges with sparse data on `hapi.fhir.org`, which future work can address by using larger FHIR datasets [14]. Charts like box plots show MIMIC-III summaries are consistent, with tight score ranges, while FHIR summaries vary more, suggesting a need for better data quality. MIMIC-III's focus on ICU patients, who are often older and have specific conditions, may bias results toward certain groups, limiting use for other patients. FHIR's standard format helps the framework scale across hospitals, enabling data sharing between systems like Epic and Cerner, but real-world integration needs testing. Security features, including AES-256 encryption and access controls, meet HIPAA and GDPR rules, making the system safe for hospitals [16, 19]. TinyLlama's small size allows it to run on regular hospital computers, unlike larger models [6]. The open-source pipeline supports further research, aligning with studies on health recommender systems. However, clinical trials with doctors are needed to confirm the system's value in real hospitals and physician feedback will help improve trust and usability.

8 Future Work

The framework will be improved to work better in hospitals. Adding data like medical images and lab results will make summaries more complete, helping doctors diagnose better. A future pilot test will involve doctors to check the system's usefulness and gather suggestions for improvement. Partnerships with other hospitals will test how well the framework works in

Table 5. Sample summary outputs.

System	MIMIC-III (10000032)	FHIR (123456)
Gold Standard	Intubation, antibiotics	Metformin 1000 mg daily
Baseline	Oxygen therapy	Insulin 10 units
MCP + TinyLlama	Intubation, antibiotics	Metformin 1000 mg daily

Table 6. Qualitative comparison of summaries.

Source	Gold Standard	Our Framework	Baseline (BioBERT)
MIMIC-III (ID: 10000032)	65yo male, CHF, treated with furosemide, discharged stable.	65yo male with CHF, managed with furosemide, stable at discharge.	Male with heart issue, treated, discharged.
FHIR (ID: 123456)	Hypertension, metformin prescribed, follow-up in 2 weeks.	Hypertension diagnosed, metformin given, follow-up in 2 weeks.	Patient with high BP, medicated.

different settings. To support global use, the system will handle non-English patient records using Noto Serif fonts for clear text display. Summaries will be made easier to understand by showing which data was used, building trust with doctors. Data from wearable devices, like heart rate from smartwatches, will be added for real-time monitoring. Tools like SHAP will be explored to explain how the system makes summaries, helping doctors understand its decisions.

9 Conclusion

This framework combines TinyLlama with MIMIC-III and FHIR data using the Model Context Protocol to create safe, shareable clinical summaries. High ROUGE-1 scores (0.9625 for MIMIC-III, 0.8500 for FHIR) and clear charts prove it works well for hospital use. The open-source pipeline at https://github.com/shekhar-ai99/clinical-mcp allows others to build on this work. Next steps include testing in a hospital sandbox, adding multimodal data like images and exploring larger models like LLaMA 3 8B for better accuracy. This system can lead the way for safe, AI-driven tools in healthcare, improving patient care worldwide.

Data Availability Statement

The source code for this study is publicly available at https://github.com/shekhar-ai99/clinical-mcp (accessed on 20 December 2025).

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

This work utilizes publicly available de-identified datasets, which do not constitute human subjects research requiring additional ethical approval.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). [CrossRef]

[3] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9. [CrossRef]

[4] Introducing the model context protocol. (n.d.). Home Anthropic. Retrieved from https://www.anthropic.com/news/model-context-protocol(accessed on 20 December 2025).

[5] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. [CrossRef]

[6] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural

language processing. *ACM Transactions on Computing for Healthcare, 3*(1), 1–23. [CrossRef]

[7] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. [CrossRef]

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020, December). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 9459-9474).

[9] Makary, M. A., & Daniel, M. (2016). Medical error—the third leading cause of death in the US. *BMJ, 353*, i2139. [CrossRef]

[10] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). [CrossRef]

[11] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out* (*WAS 2004*) (pp. 74–81).

[12] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

[13] Bender, D., & Sartipi, K. (2013, June). HL7 FHIR: An agile and RESTful approach to healthcare information exchange. In *2013 IEEE 26th International Symposium on Computer-Based Medical Systems* (pp. 326–331). IEEE. [CrossRef]

[14] Lehne, M., Sass, J., Essenwanger, A., Schepers, J., & Thun, S. (2019). Why digital medicine depends on interoperability. *NPJ Digital Medicine, 2*(1), 79. [CrossRef]

[15] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. [CrossRef]

[16] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. [CrossRef]

[17] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. [CrossRef]

[18] Hinton, G. (2018). Deep learning—A technology with the potential to transform health care. *JAMA, 320*(11), 1101–1102. [CrossRef]

[19] Li, Z., Roberts, K., Jiang, X., & Long, Q. (2019). Distributed learning from multiple EHR databases: contextual embedding models for medical events. *Journal of biomedical informatics, 92*, 103138. [CrossRef]

[20] Zhang, P., Sun, W., Li, Z., Wang, X., Li, C., Huang, R., ... & Wei, F. (2024). TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

[21] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine, 31*(3), 943-950. [CrossRef]

[22] Wang, G., Yang, G., Du, Z., Fan, L., & Li, X. (2023). ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

[23] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6), bbac409. [CrossRef]

[24] Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., ... & Oermann, E. K. (2023). Health system-scale language models are all-purpose prediction engines. *Nature, 619*(7969), 357-362. [CrossRef]

[25] Li, Y., Wang, H., Yerebakan, H. Z., Shinagawa, Y., & Luo, Y. (2024). FHIR-GPT enhances health interoperability with large language models. *NEJM AI, 1*(8), AIcs2300301. [CrossRef]

Chandra Shekhar Behera received his Master of Computer Applications degree from SRM University, Chennai, India, in 2013. He is currently pursuing his Ph.D. at Maharaja Sriram Chandra Bhanja Deo University, Baripada, India. He has over 12 years of IT experience in multinational companies, specializing in the areas of Data Science and Artificial Intelligence. His research interests include OpenAI, Agentic AI, Conversational AI, Retrieval-Augmented Generation and Large Language Models. (Email: shekhar.it99@gmail.com)

Swarupananda Bissoyi received his Ph.D. degree in Computer Science from the Department of Computer Science, Berhampur University, Berhampur, India, in 2022. He has been working as a full-time Assistant Professor at Maharaja Sriram Chandra Bhanja Deo University, Baripada, India, since 2014. Before entering academia, he gained 11 years of experience in the software industry, working for companies such as Mahindra Comviva and Samsung Research India – Bangalore. His research interests include Data Mining, Recommender Systems, Digital Humanities, Natural Language Processing, and Computer Vision. (Email: swarupananda.bissoyi@odisha.gov.in)