Privacy-Preserving Consent-Aware Retrieval-Augmented Generation System for Healthcare Queries

TEAM 7
Anukriti Singh MS2024504
Tanay Nagarkar IMT2022083

November 25, 2025

Abstract

We developed a Privacy-Preserving Consent-Aware Retrieval-Augmented Generation System for processing healthcare queries. Our system integrated privacy protection mechanisms with advanced language modeling to generate clinical reports while maintaining patient confidentiality. The pipeline processed medical queries through consent-based PII masking, semantic search using FAISS, and fine-tuned language models including Gemma-2B and Qwen2.5-1.5B. We conducted comprehensive evaluations on the user-query dataset (test set), when our train data was seen(i.e same data for RAG database and the train set for finetuning model) and unseen data (new data), with Gemma-2B achieving a BERTScore F1 of 0.87 on seen data and 0.87 on unseen data, demonstrating robust performance while ensuring data privacy through Fernet AES encryption and differential privacy approaches.

1 Introduction

In healthcare applications, generating accurate clinical reports while protecting patient privacy is crucial. We built a complete system that processes medical queries through multiple security layers and generates relevant clinical responses. Our approach combined retrieval-augmented generation with privacy-preserving techniques, allowing the system to handle sensitive medical information safely.

We focused on making the system practical for real healthcare settings by implementing consent-based data processing, where patient information could be masked or preserved based on consent flags. The entire pipeline was designed to be secure, from data ingestion through encrypted storage of final outputs.

2 Dataset Creation and Preparation

2.1 Dataset Generation Challenge

We faced a significant challenge in finding suitable healthcare datasets for our RAG pipeline. After extensive searching, we discovered that existing healthcare datasets either lacked the necessary query-report pairs required for training our system or were already

anonymized, which made them unsuitable for testing our privacy-preserving PII masking mechanisms.

To address this gap, we used DeepSeek to generate a synthetic dataset of common healthcare queries and their corresponding medical reports. This approach allowed us to create realistic clinical scenarios while maintaining control over the data structure and content.

2.2 Dataset Composition and Evaluation Strategy

We generated multiple datasets for comprehensive evaluation:

- **Training/RAG Dataset:** 478 query-report pairs used for fine-tuning and building the FAISS index
- **Seen Data Evaluation:** Test set from the same distribution as training (143 examples)
- **Unseen Data Evaluation:** Completely separate dataset of 87 query-report pairs for generalization testing
- **RAG Database:** 203 clinical reports for retrieval context

We employed the **stratified random sampling** method to ensure that both training and test sets contained representative examples of different medical conditions and query types. The training set served dual purposes: it was used to fine-tune our language models and also to build the FAISS index for our retrieval system.

3 Fine-tuning Approach

3.1 Token Length Analysis

Before fine-tuning, we conducted a comprehensive token length analysis to determine the optimal sequence length for our model. We sampled 100 examples from our dataset and built complete RAG prompts including queries, retrieved context, and instructions. The analysis revealed:

- **Minimum tokens:** 119
- **Median tokens:** 134
- **Maximum tokens:** 151
- **98th percentile:** 147 tokens

Based on this analysis, we set `MAX_LENGTH = 200` to provide adequate padding while maintaining computational efficiency.

3.2 QLoRA Configuration

We used Quantized Low-Rank Adaptation (QLoRA) for efficient fine-tuning with the following parameters:

- **Base Models:** google/gemma-2b and Qwen/Qwen2.5-1.5B
- **Quantization:** 8-bit with BitsAndBytesConfig
- **LoRA Rank (r):** 32
- **LoRA Alpha:** 64
- **LoRA Dropout:** 0.05
- **Target Modules:** `["q_proj","v_proj","k_proj","o_proj"]`

3.3 Training Configuration

The fine-tuning process used these training parameters for both models:

- **Epochs:** 20
- **Learning Rate:** 3e-4
- **Batch Size:** 1 per device with gradient accumulation steps of 8
- **Sequence Length:** 200 tokens
- **Training Strategy:** Causal language modeling with instruction masking

3.4 Model Comparison

We fine-tuned two different base models to compare their performance on clinical report generation:

- **Gemma-2B:** A specialized model from Google, fine-tuned for general language tasks
- **Qwen2.5-1.5B:** A general-purpose open-source model from Alibaba, not specifically trained for healthcare domains

Both models were fine-tuned using identical parameters, data splits, and training configurations to ensure fair comparison.

3.5 Prompt Format and Training Data

We created a structured prompt format that the models learned during fine-tuning:

```
[QUERY]
{processed_query}

[RETRIEVED_CONTEXT]
1. {retrieved_report_1}
2. {retrieved_report_2}
3. {retrieved_report_3}

[INSTRUCTION]
Generate a clinical report. Do NOT include disclaimers.
```

3.6 Privacy Protection During Inference

During the inference pipeline, we implemented consent-based PII masking for user queries:

- **60% of queries** from USER_QUERY_JSON underwent PII masking to simulate scenarios where patient consent was not granted
- **40% of queries** remained unmasked to simulate scenarios with full patient consent
- The **masked/unmasked version** was passed to both the retriever (FAISS semantic search) and the LLM (report generation)
- The **original query** was preserved in the output JSON only for evaluation and record-keeping purposes

This approach ensured that during actual processing, both the retrieval system and language model only operated on privacy-protected versions of queries when consent was not available, while maintaining the original data for performance evaluation and audit trails.

3.7 Training Process

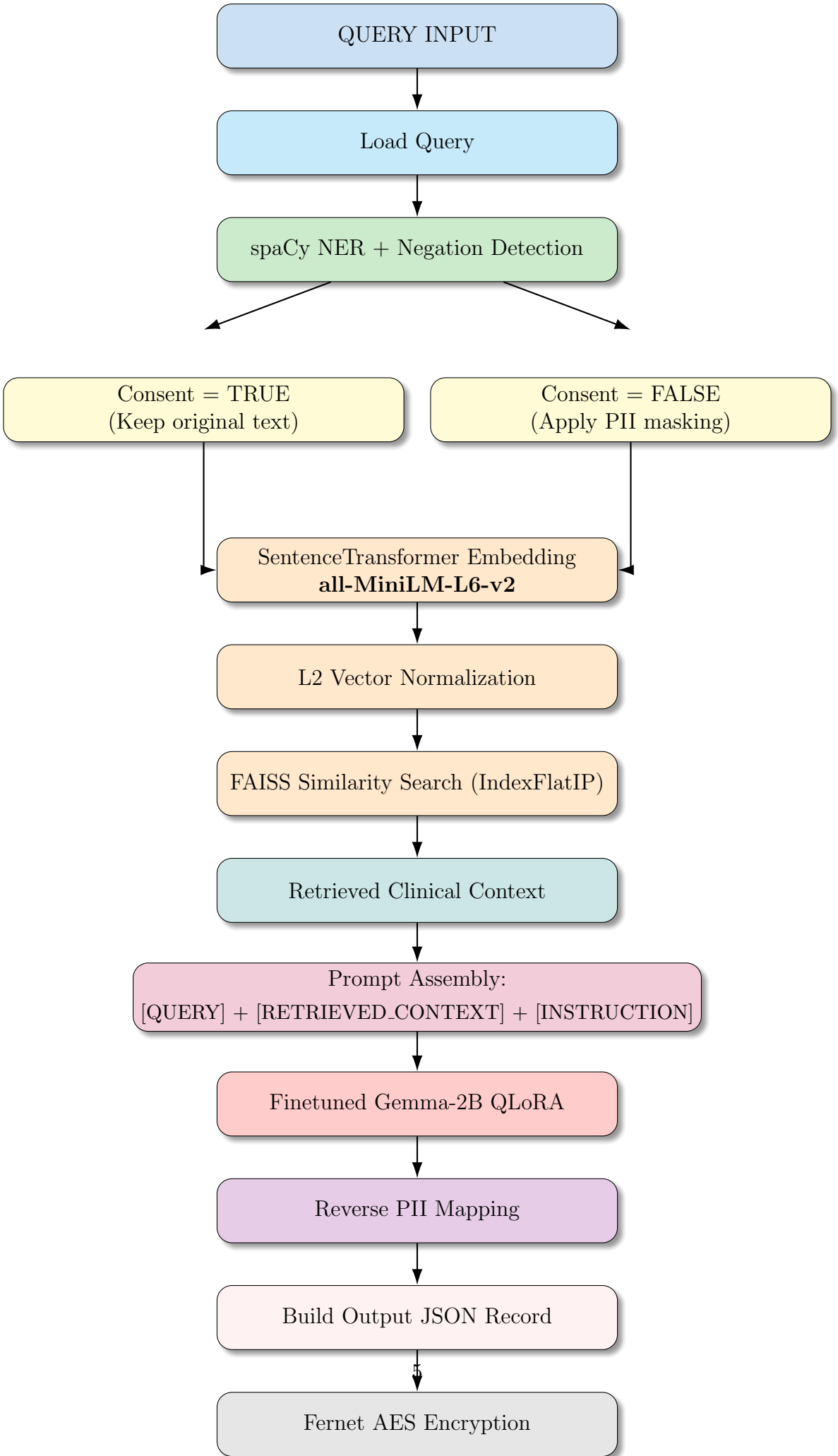
The training process involved:

1. Loading the base models with 8-bit quantization
2. Preparing the models for k-bit training
3. Adding LoRA adapters to the attention mechanisms
4. Training for 20 epochs with early stopping based on validation loss
5. Saving both the LoRA weights and the merged final models

This approach allowed us to efficiently fine-tune large language models while maintaining strong performance on clinical report generation tasks.

4 System Architecture

The complete workflow of our system followed a sequential process with specialized components at each stage:


QUERY INPUT
Load Query
spaCy NER + Negation Detection
Consent = TRUE
(Keep original text)
Consent = FALSE
(Apply PII masking)
SentenceTransformer Embedding
all-MiniLM-L6-v2
L2 Vector Normalization
FAISS Similarity Search (IndexFlatIP)
Retrieved Clinical Context
Prompt Assembly:
[QUERY] + [RETRIEVED_CONTEXT] + [INSTRUCTION]
Finetuned Gemma-2B QLoRA
Reverse PII Mapping
Build Output JSON Record
Fernet AES Encryption

5 Detailed Component Implementation

5.1 Data Processing and Privacy Protection

We started with user queries and applied multiple layers of privacy protection:

Query Input: The system accepted clinical queries in natural language format from our test dataset.

Named Entity Recognition: We used spaCy's `en_core_web_sm` model to identify entities in the text. This helped us locate potential Personally Identifiable Information (PII) that needed protection.

Negation Detection: For clinical accuracy, we implemented negation detection using the regular expression pattern: `\b(no|not|denies?|without|absent|negative for)\b`. We checked a 40-character window before each entity to determine if it was negated in the clinical context.

Consent-Based Processing: This was a core privacy feature. We randomly assigned consent flags where:

- **60% of queries** had `consent = FALSE` and underwent PII masking
- **40% of queries** had `consent = TRUE` and kept original text

5.2 PII Masking Implementation

When consent was FALSE, we applied comprehensive PII masking:

- **Names:** Used regex pattern `\b[A-Z][a-z]+ [A-Z][a-z]+\b` to find person names and replaced them with tokens like `<PHI_PERSON_1>`
- **Emails:** Pattern `[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+` identified email addresses, replaced with `<PHI_EMAIL_1>`
- **Phone Numbers:** Pattern `\b(?\d[\d\-\s]{7,}\d)\b` found phone numbers, replaced with `<PHI_PHONE_1>`
- **Dates:** Pattern `\b\d{1,2}[\-\.\s]\d{1,2}[\-\.\s]\d{2,4}\b` identified dates, replaced with `<PHI_DATE_1>`

We maintained a `mask_map` dictionary that stored all original values with their replacement tokens, allowing for reversible mapping later in the pipeline.

5.3 Semantic Search and Retrieval System

We built a robust retrieval system to find relevant clinical context:

Embedding Model: We used `all-MiniLM-L6-v2` from SentenceTransformers, which created 384-dimensional vector representations of text. This model was chosen for its balance of performance and efficiency.

FAISS Index: We used `IndexFlatIP` (Inner Product) which is optimized for cosine similarity when vectors are L2-normalized. The index contained embeddings of all clinical reports from our training set.

Retrieval Process: For each query, we:

1. Encoded the query into a 384-dimensional vector
2. Applied L2 normalization to the vector

3. Searched the FAISS index for similar documents
4. Retrieved top `TOPK=3` documents with similarity score above `SIM_THRESHOLD=0.3`
5. Used initial search of `topk*4=12` documents, then filtered by similarity threshold

5.4 Security and Storage

We implemented multiple security layers:

Encryption: Used Fernet AES encryption for all stored data. We generated encryption keys once and reused them for consistent decryption.

Secure Storage: All outputs were saved as encrypted JSON files with `.enc` extension. Each query result was stored in a separate encrypted file.

Data Handling: The original RAG data was encrypted at rest, and only decrypted temporarily during processing in memory.

6 Evaluation Methodology

We conducted comprehensive evaluations on both seen data (from the training distribution) and unseen data (completely separate dataset) to assess model performance and generalization capabilities.

	Gemma-2B		**Qwen2.5-1.5B**	
Metric	**Seen Data**	**Unseen Data**	**Seen Data**	**Unseen Data**
Retrieval Similarity	0.4686	0.4420	0.4822	0.4504
Semantic Similarity	0.5482	0.5452	0.5472	0.5605
Faithfulness	0.5538	0.5297	0.5779	0.5104
Answer Relevancy	0.5089	0.5110	0.5094	0.5125
BERTScore F1	0.8719	0.8651	0.8616	0.8392
ROUGE-L	0.2530	0.2504	0.2603	0.2497
BLEU	0.0978	0.0952	0.1084	0.0862

Table 1: Comprehensive Performance Evaluation on Seen and Unseen Data

Evaluation Strategy: We tested both models under two conditions:

- **Seen Data:** 143 examples from the same distribution as training data
- **Unseen Data:** 87 completely separate query-report pairs for generalization testing
- **RAG Database:** 203 clinical reports used as retrieval context

Model Comparison Approach: We fine-tuned Qwen2.5-1.5B, a non-healthcare open-source model, using the exact same methods and parameter values as our primary Gemma-2B model. This allowed us to compare how a general-purpose model would perform on healthcare-specific tasks when fine-tuned with our RAG pipeline.

Key Performance Insights:

- **Gemma-2B** showed consistent performance across both seen and unseen data, with BERTScore F1 maintaining at 0.87, demonstrating strong generalization capability.

- **Qwen2.5-1.5B** performed competitively on seen data but showed a noticeable drop in BERTScore F1 (from 0.86 to 0.84) on unseen data, suggesting slightly lower generalization.
- **Retrieval Performance** was similar across both models, with Qwen showing slightly better retrieval similarity on seen data but both models experiencing expected drops on unseen data.
- **Semantic Understanding** remained strong for both models, with Qwen actually showing improved semantic similarity on unseen data (0.5605 vs 0.5472).
- **Faithfulness** metrics indicate how well generated answers align with retrieved documents, with both models maintaining reasonable alignment.

7 Key Technical Decisions

7.1 Privacy-First Approach

We prioritized patient privacy through multiple mechanisms:

- **Consent-based processing** with 60% masking rate during training
- **Reversible PII mapping** that only restored information when consent was TRUE
- **End-to-end encryption** for all stored data
- **Secure key management** with persistent Fernet keys

7.2 Clinical Domain Adaptation

We tailored the system for healthcare:

- **Negation detection** for accurate clinical understanding
- **Biomedical entity awareness** in retrieval scoring
- **Structured prompt format** that matched clinical reporting style
- **No-disclaimer instruction** to avoid unnecessary legal text in outputs

7.3 Optimizations we did

We balanced performance with computational efficiency:

- **8-bit quantization** for manageable memory usage
- **QLoRA fine-tuning** to avoid full model retraining
- **FAISS indexing** for fast similarity search
- `all-MiniLM-L6-v2` for efficient embeddings without sacrificing quality

8 Conclusion

We successfully developed a Privacy-Preserving Consent-Aware Retrieval-Augmented Generation System that processes healthcare queries while maintaining strong privacy protections. Our comprehensive evaluation demonstrates that both Gemma-2B and Qwen2.5-1.5B can be effectively fine-tuned for clinical report generation tasks.

Key Findings:

- Gemma-2B showed excellent consistency between seen and unseen data, with BERTScore F1 maintaining at 0.87, indicating strong generalization capability.
- Qwen2.5-1.5B performed competitively on seen data but showed some performance degradation on unseen data, suggesting that specialized models like Gemma may have advantages in healthcare domains.
- Both models maintained reasonable performance across all evaluation metrics, demonstrating the effectiveness of our RAG pipeline approach.
- The privacy-preserving mechanisms successfully protected sensitive information while maintaining system utility.

The modular architecture allows for future improvements in individual components, such as more sophisticated entity recognition or larger language models. The consent-based processing framework provides a foundation for compliance with healthcare regulations like HIPAA, while the comprehensive evaluation metrics ensure ongoing quality monitoring.

This work shows that with careful design choices, it's possible to build AI systems for healthcare that balance utility with privacy, opening possibilities for safer adoption of AI in medical settings. The comparative analysis provides valuable insights for practitioners selecting models for healthcare applications.

References

[1] *Privacy-Preserving Retrieval-Augmented Generation with Differential Privacy*, arXiv:2412.04697, 2024. Available: `https://arxiv.org/abs/2412.04697`

[2] *Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval*, arXiv:2503.15548, 2025. Available: `https://arxiv.org/abs/2503.15548`

[3] Tim Dettmers et al., *QLoRA: Efficient Finetuning of Quantized LLMs*, arXiv:2305.14314, 2023.

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou, *Billion-scale similarity search with GPUs*, IEEE Transactions on Big Data, 2019.

[5] W. Wang et al., *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Embeddings*, arXiv:2002.10957, 2020.

[6] Matthew Honnibal et al., *spaCy: Industrial-strength NLP in Python*, Explosion AI, 2020.

[7] Google DeepMind, *Gemma: Open Models Based on Gemini Research and Technology*, arXiv:2403.08295, 2024.

[8] Alibaba Qwen Team, *Qwen2.5 Technical Report*, arXiv:2409.12191, 2024.

[9] D. DiBari, *Fernet Specification*, Cryptography.io Documentation, 2018.

[10] C. Dwork, *Differential Privacy: A Survey of Results*, TAMC, 2008.