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

TEAM 7
Anukriti Singh MS2024504
Tanay Nagarkar IMT2022083

Abstract

We developed a Privacy-Preserving Consent-Aware Retrieval-Augmented Generation System for healthcare queries. The system integrates privacy protection with language modeling to generate clinical reports while maintaining patient confidentiality. The pipeline processes medical queries through consent-based PII masking, semantic search using FAISS, and fine-tuned language models including Gemma-2B and Qwen2.5-1.5B. Evaluations on both seen and unseen data show Qwen2.5-1.5B achieved BERTScore F1 of 0.88 on seen data and 0.86 on unseen data, with consistent 60% PII masking and Fernet encryption throughout the pipeline.

1 Introduction

In healthcare applications, generating accurate clinical reports while protecting patient privacy presents unique challenges. Using LLMs alone creates risk of providing outdated or unverified information, which is particularly dangerous in medical contexts. Our system addresses this by combining Retrieval-Augmented Generation (RAG) with privacy-preserving techniques which ensures responses are grounded in verified medical knowledge by retrieving relevant clinical (user,report)data.Our consent-based processing where patient information is masked based on consent flags and the encryption of the generated outputs creates a complete pipeline that maintains privacy of user data and protection against external attacks.

2 Dataset Creation and Preparation

2.1 Dataset Generation Challenge

Available healthcare datasets lacked suitable query-report pairs or were already anonymized, making them unsuitable for testing our privacy-preserving PII masking. We used DeepSeek to generate a synthetic dataset of healthcare queries and corresponding medical reports, allowing control over data structure and content.

2.2 Dataset Composition and Evaluation Strategy

We generated datasets for evaluation:

- Training/RAG Dataset: Query-report pairs for fine-tuning and FAISS index

- Seen Data Evaluation: Test set from training distribution
- Unseen Data Evaluation: Separate dataset for generalization testing
- RAG Database: Clinical reports for retrieval context

We used 80-20 split with 60% PII masking applied to both training and test sets. The training set was used for both fine-tuning and building the FAISS index.

2.3 Consistent Masking Strategy

We implemented consistent masking:

- 60% masking probability across all datasets
- Both training and test splits received identical masking
- PII categories: PERSON, GPE, LOC
- Regex patterns for names, emails, and phone numbers
- Original queries preserved for evaluation

3 Fine-tuning Approach

3.1 Token Length Analysis

Analysis of RAG prompts showed 200 tokens provided adequate padding while maintaining efficiency.

3.2 QLoRA Configuration

We used QLoRA for efficient fine-tuning:

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

3.3 Training Configuration

Training parameters for both models:

- Epochs: 20
- Learning Rate: 2e-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 models:

- Gemma-2B: Specialized model from Google
- Qwen2.5-1.5B: General-purpose model from Alibaba

Both used identical parameters and configurations.

3.5 Prompt Format and Training Data

Structured prompt format:

```
[QUERY]
{processed_query}

[SIMILAR_CASES]
Case 1:
Query: {similar_query_1}
Report: {report_1}
...
[INSTRUCTION]
Generate a clinical report. Do NOT include disclaimers.
```

3.6 Privacy Protection During Training and Inference

Privacy protection included:

- 60% of queries underwent PII masking
- 40% of queries remained unmasked
- Masked/unmasked versions passed to retriever and LLM
- Original query preserved for evaluation
- Additional masking for data augmentation

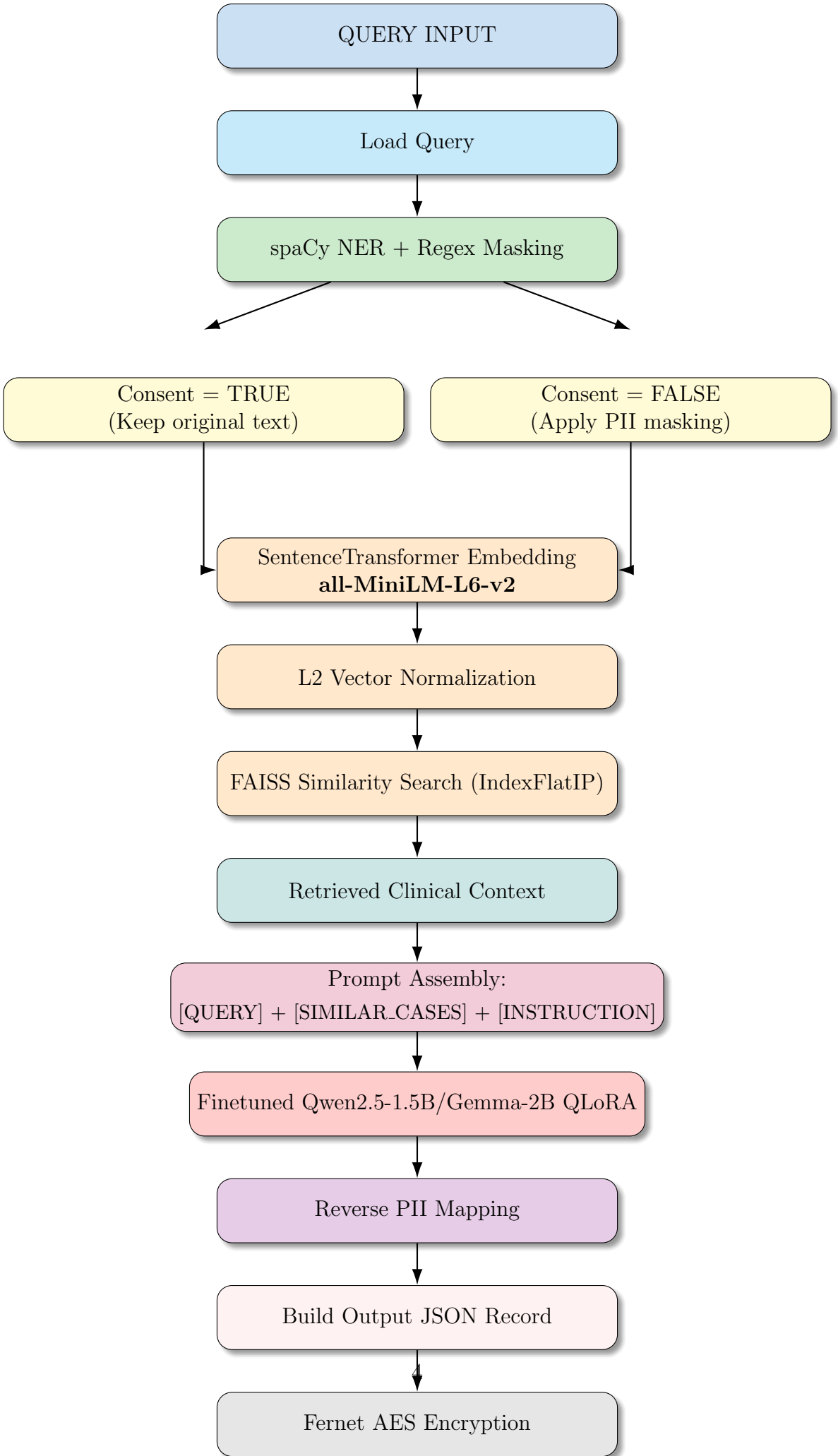
3.7 Training Process

The training process:

1. Load base models with 8-bit quantization
2. Prepare for k-bit training
3. Add LoRA adapters
4. Train for 20 epochs with gradient clipping
5. Save LoRA weights and merged models

4 System Architecture

The system workflow:


QUERY INPUT
Load Query
spaCy NER + Regex Masking
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] + [SIMILAR_CASES] + [INSTRUCTION]
Finetuned Qwen2.5-1.5B/Gemma-2B QLoRA
Reverse PII Mapping
Build Output JSON Record
Fernet AES Encryption

5 Detailed Component Implementation

5.1 Data Processing and Privacy Protection

The system processes queries through multiple privacy layers:
Query Input: Clinical queries in natural language format.
Named Entity Recognition: spaCy's `en_core_web_sm` model to identify PII.
Consent-Based Processing: Random consent flags:

- 60% of queries: `consent = FALSE` with PII masking
- 40% of queries: `consent = TRUE` with original text

5.2 PII Masking Implementation

When consent was FALSE:

- Names: Regex pattern with `[NAME]` replacement
- Emails: Pattern with `[EMAIL]` replacement
- Phone Numbers: Pattern with `[PHONE]` replacement
- Entity-based: spaCy labels with tokens like `<PERSON_1>`

A `mask_map` dictionary stored original values for reversible mapping.

5.3 Semantic Search and Retrieval System

Retrieval system components:
Embedding Model: `all-MiniLM-L6-v2` for 384-dimensional vectors.
FAISS Index: `IndexFlatIP` optimized for cosine similarity.
Retrieval Process:

1. Encode query to vector
2. Apply L2 normalization
3. Search FAISS index
4. Retrieve top 3 documents above 0.3 similarity threshold
5. Initial search of 12 documents, then filter

5.4 Security and Storage

Security measures:

- Fernet AES encryption for stored data
- Encrypted JSON files with `.enc` extension
- Separate encrypted files for each query result
- Data encrypted at rest, decrypted only during processing

	Qwen2.5-1.5B		Gemma-2B	
Metric	**Seen Data**	**Unseen Data**	**Seen Data**	**Unseen Data**
Retrieval Similarity	0.3714	0.3242	0.3688	0.3227
Semantic Similarity	0.6040	0.5967	0.5168	0.4857
Faithfulness	0.4155	0.3716	0.3747	0.3541
Answer Relevancy	0.5813	0.5076	0.5312	0.4654
BERTScore F1	0.8802	0.8619	0.8623	0.8558
ROUGE-L	0.2783	0.2971	0.2268	0.1994
BLEU	0.0911	0.1137	0.0758	0.0724

Table 1: Performance Evaluation on Seen and Unseen Data

6 Evaluation Methodology

Evaluations on seen data (training distribution) and unseen data (separate dataset).
Evaluation Strategy:

- Seen Data: Examples from training distribution
- Unseen Data: Separate query-report pairs

Model Comparison: Both models fine-tuned with same methods and parameters.
Performance Insights:

Based on the evaluation metrics, Qwen achieved higher scores than Gemma across most scores. For Retrieval Similarity (document relevance matching), Qwen scored 0.3714 on seen data and 0.3242 on unseen data compared to Gemma's 0.3688 and 0.3227. For Semantic Similarity (meaning preservation), Qwen reached 0.6040 and 0.5967 versus Gemma's 0.5168 and 0.4857. In Faithfulness (alignment with source content), Qwen achieved 0.4155 and 0.3716 while Gemma scored 0.3747 and 0.3541. For Answer Relevancy (response appropriateness), Qwen obtained 0.5813 and 0.5076 compared to Gemma's 0.5312 and 0.4654. In text generation quality, Qwen's BERTScore F1 (semantic similarity) was 0.8802 and 0.8619 while Gemma's was 0.8623 and 0.8558. For ROUGE-L (longest common subsequence), Qwen scored 0.2783 and 0.2971 versus Gemma's 0.2268 and 0.1994. In BLEU (n-gram precision), Qwen achieved 0.0911 and 0.1137 compared to Gemma's 0.0758 and 0.0724. Performance decreased on unseen data for both models across most metrics, while Qwen maintaining slightly better scores.

7 Key Technical Decisions

7.1 Privacy-First Approach

Privacy mechanisms:

- Consent-based processing with 60% masking
- Reversible PII mapping
- End-to-end encryption
- Secure key management

7.2 Clinical Domain Adaptation

- Structured prompt format with clinical context
- PII-aware processing for healthcare entities
- No-disclaimer instruction
- Evaluation on seen and unseen medical data

8 Conclusion

We developed a Privacy-Preserving Consent-Aware RAG System for healthcare queries. Evaluation shows both Qwen2.5-1.5B and Gemma-2B can be effectively fine-tuned for clinical report generation.

Key Findings:

- Qwen2.5-1.5B: BERTScore F1 0.8802 on seen, 0.8619 on unseen data
- Gemma-2B: BERTScore F1 0.8623 on seen, 0.8558 on unseen data
- Qwen outperformed Gemma in semantic similarity and text generation
- Both models maintained reasonable performance across metrics
- Privacy mechanisms protected sensitive information

The modular architecture allows component improvements. The consent-based framework supports healthcare regulation compliance. This work demonstrates AI systems for healthcare can balance utility with privacy.

References

[1] *Privacy-Preserving Retrieval-Augmented Generation with Differential Privacy*, arXiv:2412.04697, 2024.

[2] *Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval*, arXiv:2503.15548, 2025.

[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.