# **Biomedical RAG System Development Report**

**Date**: August 19, 2025

## **Executive Summary**

This report presents the development and evolution of a Retrieval-Augmented Generation (RAG) system specifically designed for biomedical question answering. Through six iterative versions, the system evolved from a basic prototype to a sophisticated, domain-optimized solution capable of processing scientific literature and providing accurate, grounded responses.

## **Project Overview**

**Objective**: Develop a biomedical question-answering system that can retrieve relevant information from scientific papers and generate accurate, contextually grounded responses.

**Final Achievement**: Successfully deployed a functional RAG system with a user-friendly Gradio interface, accessible via web deployment.

## **Dataset & Knowledge Base**

**Data Sources**: Curated collection of 13 peer-reviewed biomedical research papers focusing on AI-driven drug discovery and computational biology

### **Paper Collection Overview**

**Primary Focus Areas**:

* **Drug Discovery & AI** (5 papers): AI applications, generative models, and computational methods
* **Biomedical Data Mining** (4 papers): Patent mining, bioactivity databases, NLP for target extraction
* **GPCR Research** (3 papers): Structural insights, ligand interactions, and drug design
* **Retrieval-Augmented Systems** (2 papers): RAG applications in biomedicine and clinical Q&A

### **Specific Research Papers**

1. **Artificial Intelligence in Drug Discovery**: Applications, challenges, and future directions
2. **Patent Bioactivity Mining**: SureChEMBL and IDG for understudied drug targets
3. **Cheminformatics Tools**: Public bioactivity databases in drug discovery
4. **Biomedical NLP**: Target and relationship extraction from literature
5. **Deep Generative Models**: De novo drug design methods and applications
6. **RAG in Biomedicine**: Systematic review of retrieval-augmented generation
7. **Clinical RAG Meta-Analysis**: Large language models for biomedical Q&A
8. **GPCRdb 2025**: Structure and bioactivity data integration for GPCR drug design
9. **IFMoAP**: Multimodal mechanism-of-action prediction using cell painting
10. **Rag2Mol**: Structure-based drug design with pocket-specific fragments
11. **Morphological Profiling**: Cell painting for drug mechanism and repurposing
12. **DeepGPCR**: Graph neural networks for GPCR-ligand interaction prediction
13. **M5 Muscarinic Receptor**: Structural insights via cryo-EM analysis

**Data Processing Pipeline**:

* PDF text extraction with formatting artifact removal
* Category-aware chunking based on paper classifications (overview, methods, resources, structure)
* Section-aware segmentation to preserve document structure and context
* Quality filtering and preprocessing for optimal retrieval performance

## **Development Evolution**

### **Version 1: Initial Prototype**

* **Corpus**: 5 biomedical papers
* **Embedding Model**: all-MiniLM-L6-v2 (general-purpose)
* **Answer Generation**: BioGPT-Large
* **Challenges**: Messy, repetitive outputs due to long input chunks and PDF formatting artifacts
* **Innovation**: Implemented LLM-as-Judge validation and biomedical summarization (venkyb/bart-finetuned-summarization-pubmed)

### **Version 2: Domain Optimization**

* **Key Improvements**:
  + Upgraded to BioBERT SentenceTransformer (pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb)
  + Implemented coherent chunking (800 characters with sentence boundaries)
  + Added FAISS distance threshold filtering
* **Results**: Significantly improved relevance and accuracy through domain-specific embeddings

### **Version 3: Scalability Enhancement**

* **Major Advances**:
  + Expanded corpus to 13 research papers
  + Implemented section-aware chunking with overlapping segments
  + Upgraded to allenai-specter embedding model (scientific literature-trained)
  + Integrated BioMistral-7B for answer generation with source citations
* **Limitation**: Required GPU resources not available on development system

### **Version 4: CPU Optimization**

* **Focus**: Efficient CPU execution with Flask web API
* **Architecture**: Thread pool for parallel request handling with unique task IDs
* **Models**: allenai-specter embeddings + lightweight BioLLaMA
* **Issue**: Answer generation configuration problems led to raw chunk outputs

### **Version 5: Interface Integration**

* **Platform**: Streamlit-based user interface
* **Model**: BioMistral-7B-GGUF via ctransformers
* **Optimization**: CPU-optimized with strict context limits
* **Challenge**: Prompt length exceeded model's maximum context capacity

### **Version 6: Production Solution**

* **Final Architecture**:
  + **Embedding**: S-PubMedBert-MS-MARCO (domain-specific)
  + **Retrieval**: FAISS with efficient similarity search
  + **Generation**: Llama 2 via Ollama API
  + **Interface**: Gradio web application
* **Deployment**: Successfully deployed with public access
* **URL**: <https://e73aa738dbc97991e0.gradio.live/>

## **Technical Achievements**

### **Core Innovations**

1. **Domain-Specific Optimization**: Progression from general-purpose to biomedical-specialized models
2. **Advanced Chunking**: Evolution from simple text splitting to section-aware, overlapping segments
3. **Robust Retrieval**: Implementation of FAISS indexing with distance threshold filtering
4. **Context Management**: Dynamic context packing to optimize LLM input utilization
5. **Source Transparency**: Citation system for answer provenance

### **System Architecture**

* **Embedding Layer**: S-PubMedBert-MS-MARCO for semantic understanding
* **Storage & Retrieval**: FAISS vector database for efficient similarity search
* **Generation**: Llama 2 integration via Ollama API
* **Interface**: User-friendly Gradio web application
* **Deployment**: Cloud-accessible solution

## **Challenges Overcome**

1. **PDF Processing**: Resolved formatting artifacts and noise in extracted text
2. **Context Length Limitations**: Implemented dynamic chunking and context management
3. **Resource Constraints**: Adapted system for CPU-only execution
4. **Output Quality**: Eliminated repetitive and incoherent responses through iterative refinement
5. **Scalability**: Successfully expanded from 5 to 13+ research papers

## **Business Impact**

### **Immediate Benefits**

* **Accessibility**: Web-deployed solution available for immediate use
* **Efficiency**: Automated biomedical literature search and synthesis
* **Accuracy**: Domain-optimized models ensure relevant, grounded responses
* **Transparency**: Source citation enables verification and further research

### **Potential Applications**

* **Drug Discovery Acceleration**: Query-based access to AI methods, generative models, and computational approaches
* **Target Identification**: Rapid literature search for understudied proteins and bioactive compounds
* **GPCR Research Support**: Specialized knowledge base for G-protein coupled receptor studies
* **Methodology Discovery**: Quick access to experimental methods, databases, and analytical tools
* **Clinical Decision Support**: Evidence-based responses from curated biomedical literature

## **Lessons Learned**

1. **Domain Specialization**: Biomedical-specific models significantly outperform general-purpose alternatives
2. **Iterative Development**: Each version addressed specific limitations while building on previous successes
3. **Resource Planning**: Early consideration of computational requirements prevents deployment bottlenecks
4. **User Experience**: Interface design is crucial for adoption and practical utility

## **Conclusion**

The biomedical RAG system represents a successful evolution through six development iterations, culminating in a production-ready solution. The system demonstrates significant technical achievements in domain specialization, efficient retrieval, and user accessibility. The deployed solution provides immediate value while establishing a foundation for future enhancements and broader applications in biomedical research and clinical support.

**Status**: Successfully Deployed and Operational

**Access**: <https://e73aa738dbc97991e0.gradio.live/>