Archives available at journals.mriindia.com

**Open Access International Journal of Science and Engineering**

ISSN: 2456-3293

Volume 9 Issue 03, 2026

# Offline Biomedical Knowledge Engine for Evidence-Grounded Summarization Using Medingest AI

[1]Samidha Arun Kasare , [2]Dr. Ankita Karale, [3]Dr. Balkrishna K. Pati, [4]Dr. Naresh Thoutam
*[1]Student, Department of Computer Engineering, SITRC, Nashik-422213, India*
*[2,3,4]Dr. Department of Computer Engineering, SITRC, Nashik-422213, India[2]*
*Email: [1]kasaresamidha67@gmail.com, [2]ankita.karale@sitrc.org, [3]balkrishnapatileng@gmail.com, [4]naresh.thoutam@sitrc.org*

**Peer Review Information**

*Submission: 10 Feb 2026*

*Revision: 26 Feb 2026*

*Acceptance: 11 March 2026*

**Keywords**

*Biomedical Knowledge Engine, Retrieval-Augmented Generation, Offline AI System, Evidence-Based Summarization, MedIngest AI, Document Ingestion, Semantic Search, Citation-Based Summaries, Natural Language Processing, Biomedical Information Retrieval*

**Abstract**

The rapid growth of biomedical research data has made it difficult for researchers and clinicians to quickly access reliable and relevant information. Manual analysis of large collections of research papers is time-consuming and often leads to missed insights. Although modern AI-based systems can generate summaries, they frequently lack factual grounding and depend heavily on internet-based services. This work presents MedIngest AI, an offline biomedical knowledge engine designed to ingest PDF documents, retrieve relevant evidence, and generate citation-backed summaries. The system follows a structured pipeline that includes text extraction, chunking, embedding generation, indexing, retrieval, and summarization. A key focus is placed on maintaining evidence linkage and reliability throughout the process. The system is implemented as a modular platform with an interactive user interface that supports document management, query-based summarization, analytics visualization, and export functionality. Screens such as the document library, summarization interface, analytics dashboard, and export center demonstrate the practical usability of the system. Overall, the proposed system provides a self-contained and privacy-preserving solution for biomedical knowledge processing. It improves transparency by linking generated summaries with source evidence and enables users to work efficiently without relying on external cloud services.

## Introduction

Biomedical research has grown rapidly over the years, leading to a massive increase in scientific publications, clinical reports, and datasets. While this expansion improves access to knowledge, it also creates a challenge for researchers and clinicians who need to quickly extract useful information from large volumes of text. Manual reading and analysis are no longer practical, especially when dealing with multiple documents and complex medical content. [4]

At the same time, artificial intelligence and natural language processing have introduced automated ways to summarize and analyze text. These systems can generate summaries in a short time, but they often lack reliability. Many existing solutions produce content that is not fully supported by source documents, which reduces trust in the output. In biomedical applications, even small inaccuracies can lead to serious

consequences, making accuracy and traceability very important. [17]

Another limitation of current systems is their dependence on cloud-based services. Most tools require internet connectivity and external APIs to process data. This raises concerns about data privacy, especially when handling sensitive medical information. It also limits usability in environments where internet access is restricted or not allowed. [6]

To address these challenges, this work introduces MedIngest AI, an offline biomedical knowledge engine designed to process documents, retrieve relevant information, and generate summaries with proper evidence support. The system focuses on providing a reliable and transparent workflow where each generated output can be traced back to its source. Instead of relying only on text generation, it combines retrieval, processing, and validation steps to improve the quality of results. [10]

The proposed system also emphasizes usability through an interactive interface. Users can upload documents, explore the document library, generate summaries, and analyze system performance using built-in dashboards. By combining system design with practical implementation, the platform provides a complete solution for biomedical knowledge processing in offline environments.

**Objectives Of the System**

The primary goal of MedIngest AI is to build a reliable and fully offline system that can process biomedical documents and generate evidence-backed summaries. The system is designed to address both technical and practical challenges in biomedical text processing by focusing on accuracy, transparency, and usability.

The key objectives of the system are as follows:

- Offline Document Processing:
  To develop a system that can ingest and process biomedical PDF documents without requiring internet connectivity, ensuring data privacy and secure execution.
- Efficient Text Extraction and Indexing:
  To extract meaningful textual content from documents, divide it into structured chunks, and store it in an indexed format for fast retrieval.
- Evidence-Based Retrieval Mechanism:
  To implement a retrieval process that identifies the most relevant document segments based on user queries using semantic similarity techniques.
- Citation-Backed Summarization:
  To generate summaries that are directly linked to source documents, ensuring that every statement can be traced to its original evidence.
- Reliability and Transparency:
  To maintain trust in the system by providing evidence linkage and ensuring that outputs are grounded in retrieved data.
- User-Friendly Interface:
  To design an interactive interface that allows users to upload documents, query the system, view results, and explore data easily.
- Analytics and Monitoring:
  To provide system-level insights such as number of documents, processed chunks, and index size for better understanding of system performance.
- Export Functionality:
  To enable users to export generated summaries in formats such as DOCX and Markdown for further use in research or reporting.

These objectives ensure that the system not only performs technical tasks effectively but also remains practical for real-world biomedical applications where reliability, usability, and data security are critical.

**System Overview**

MedIngest AI is designed as a complete offline platform that transforms raw biomedical documents into structured, queryable knowledge. The system integrates document ingestion, semantic retrieval, summarization, and visualization into a single workflow, allowing users to interact with biomedical data in a simple and efficient way. [5]

At the core, the system begins with document ingestion, where users upload biomedical PDF files. These documents are processed to extract textual content, which is then divided into smaller chunks for better handling. Each chunk is converted into vector representations and stored in a local index, enabling fast and accurate retrieval during query processing.

Once the data is prepared, the system allows users to enter queries through an interactive interface. The retrieval component searches the indexed data and selects the most relevant chunks based on semantic similarity. These retrieved segments act as the foundation for generating summaries, ensuring that the output

remains grounded in actual document content. [10]

The summarization module then processes the retrieved information and produces a concise response. Unlike generic summarization tools, the system maintains a strong connection between the generated content and the original documents. This helps in improving transparency and allows users to verify the information easily.

In addition to summarization, the system includes supporting components such as analytics and export modules. The analytics section provides insights into system activity, including the number of documents processed, chunks created, and index size. The export feature enables users to save generated summaries in structured formats for further use. [2]

The entire workflow is managed through a user interface that connects all components seamlessly. From document upload to final output, each step is designed to be intuitive and transparent. This makes the system suitable for researchers, students, and professionals who need a reliable tool for biomedical knowledge extraction without relying on external services.

Overall, the system provides a unified solution where ingestion, retrieval, summarization, and visualization work together to deliver accurate and evidence-based results in an offline environment.

**System Architecture**

The system architecture of MedIngest AI is designed as a modular and layered framework where each component performs a specific task in the overall pipeline. The architecture ensures smooth data flow from document ingestion to final summary generation while maintaining reliability, traceability, and offline execution.

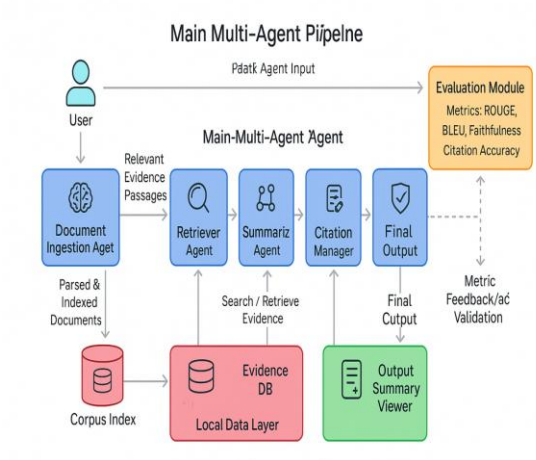


*Figure 1: System Architecture*

At a high level, the architecture is divided into multiple functional layers:

**Input and Ingestion Layer**

This layer handles the entry of biomedical documents into the system. Users upload PDF files through the interface, and the ingestion module processes these files by extracting text content. The extracted data is cleaned, normalized, and divided into smaller chunks to make it suitable for further processing.

This step is important because large biomedical documents are complex, and breaking them into smaller segments improves retrieval accuracy and system performance.

**Indexing and Storage Layer**

Once the documents are processed, the system converts each text chunk into vector embeddings. These embeddings are stored in a local index, which acts as the core knowledge base of the system. The indexing mechanism allows fast similarity-based search during query execution.

All data is stored locally, ensuring privacy and eliminating the need for external storage or cloud services.

**Retrieval Layer**

The retrieval module is responsible for selecting the most relevant document segments based on the user's query. It uses semantic similarity techniques to match the query with indexed embeddings and returns the top relevant chunks.

This layer ensures that only meaningful and contextually relevant information is passed to the next stage, which improves the quality of generated summaries.

**Summarization Layer**

In this layer, the system generates a concise summary using the retrieved content. The summarization process is controlled to ensure that the output remains aligned with the input evidence. The system avoids introducing unrelated information by restricting generation to retrieved data.

This approach improves the reliability and consistency of the output.

**Evidence and Citation Layer**

After generating the summary, the system links the output to its corresponding source documents. This ensures that users can trace each piece of information back to its origin.

This layer plays a key role in maintaining transparency and trust, especially in biomedical applications where evidence validation is critical.

**Analytics and Monitoring Layer**

The system includes an analytics module that provides insights into internal operations. It tracks parameters such as number of uploaded documents, processed chunks, and index size.

This helps users understand system performance and monitor data processing activities effectively.

**Output and User Interface Layer**

The final layer is responsible for presenting results to the user. The interface displays summaries, retrieved content, and system analytics in a structured manner. It also provides options for exporting results into different formats.

The interface connects all modules and ensures a smooth user experience from input to output.

**Architectural Advantages**

- Modular design ensures easy maintenance and scalability
- Offline operation guarantees data privacy and reproducibility
- Evidence-linked output improves transparency and trust
- Efficient retrieval enhances system performance
- Integrated UI improves usability and accessibility

Overall, the architecture ensures that each component works independently while contributing to a cohesive and reliable system for biomedical knowledge processing.

**Methodology**

The working of MedIngest AI follows a structured pipeline where each step transforms raw biomedical documents into meaningful and evidence-supported summaries. The workflow is designed to be sequential, ensuring that data flows through well-defined stages without loss of information or accuracy.

**Step-by-Step Workflow**

The complete workflow of the system is described below:

1. Document Upload
   The process begins when the user uploads biomedical PDF documents through the interface. These documents form the input dataset for the system.
2. Text Extraction and Cleaning
   The system extracts textual content from the uploaded PDFs. It removes unnecessary elements and normalizes the data to ensure consistency.
3. Chunking of Documents
   The extracted text is divided into smaller chunks. This improves the efficiency of indexing and retrieval by allowing the system to work with manageable text units.
4. Embedding Generation
   Each text chunk is converted into a vector representation. These embeddings capture the semantic meaning of the text and are used for similarity-based retrieval.
5. Index Creation
   The generated embeddings are stored in a local index. This index acts as the core storage system for all processed data and enables fast search operations.
6. Query Input
   The user provides a query through the system interface. This query represents the information requirement.
7. Semantic Retrieval
   The system searches the index to find the most relevant text chunks based on the query. Only the top matching results are selected.
8. Summary Generation
   The retrieved chunks are passed to the summarization module, which generates a concise and meaningful summary.
9. Evidence Linking
   The generated summary is associated with its source content, allowing users to verify the information.
10. Result Display and Export
    The final output is displayed through the interface, and users can export the summary for further use.

**Workflow Characteristics**

- The process is fully offline, ensuring secure data handling
- Each stage is dependent on the previous step, maintaining data consistency
- Retrieval-based processing ensures relevance of results
- Evidence linkage improves transparency and trust

**Logical Flow Representation**

The workflow can be summarized as:

Document Upload → Text Processing → Chunking → Embedding → Indexing → Query → Retrieval → Summarization → Evidence Linking → Output

This linear flow ensures that the system remains organized and predictable, which is important for both debugging and scalability.

**Importance of Workflow Design**

The structured workflow is essential for maintaining the reliability of the system. Instead of generating summaries directly from raw input, the system first filters and retrieves relevant information. This reduces noise and improves the quality of the final output.

By combining multiple steps such as retrieval, processing, and summarization, the system ensures that the output is both meaningful and grounded in actual data.

This methodology provides a clear and efficient way to process biomedical documents, making the system suitable for real-world applications where accuracy and usability are equally important.

## System Implementation

The implementation of MedIngest AI focuses on providing a practical and interactive platform where users can upload documents, explore data, generate summaries, and export results. The system is designed with a user-friendly interface that connects all backend modules into a single workflow. The following section explains the implementation using system screenshots to demonstrate functionality and usability.

### Dashboard Overview

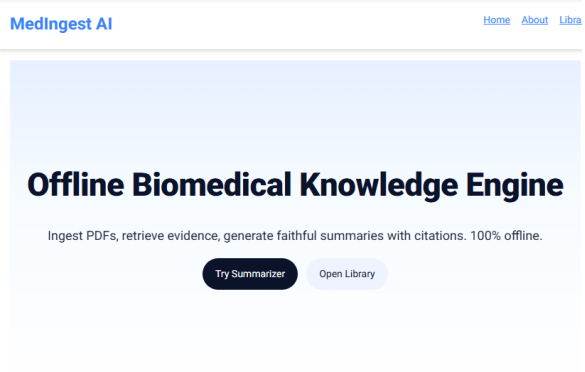


*Figure 2: Main Dashboard Interface*

This screen represents the central dashboard of the system. It displays key system statistics such as total documents, processed chunks, index size, and system status. The dashboard provides a quick overview of the system's current state and helps users monitor activity efficiently.

### Document Library Module

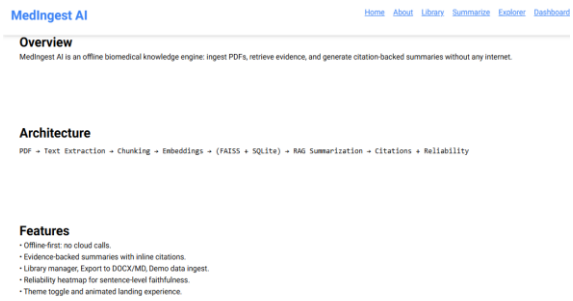


*Figure 3: Document Library and Upload Interface*

This interface allows users to upload biomedical PDF documents and manage stored files. Users can view document names, sizes, and upload timestamps. This module is important because it acts as the entry point for data ingestion and ensures that all documents are organized systematically.

### Summarization Interface

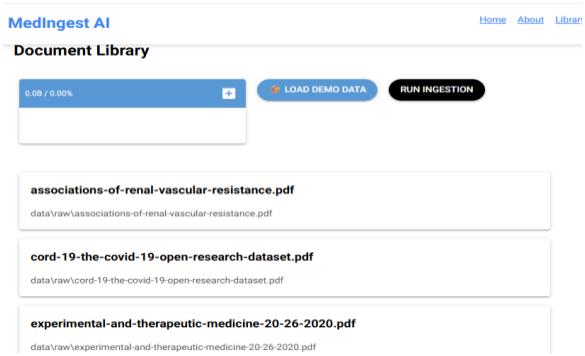


*Figure 4: Query-Based Summarization Screen*

In this screen, users enter queries related to biomedical topics. The system retrieves relevant content and generates a summary based on the indexed data. This interface is the core interaction point where users obtain insights from uploaded documents.

### Generated Summary Output

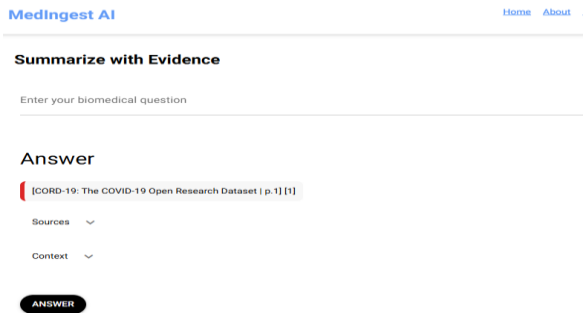


*Figure 5: Generated Summary with Evidence*

This screen shows the output summary generated by the system. The content is derived from retrieved document segments and reflects the system's ability to produce concise and meaningful responses. The output demonstrates how information is structured for easy understanding.

**Analytics Dashboard**

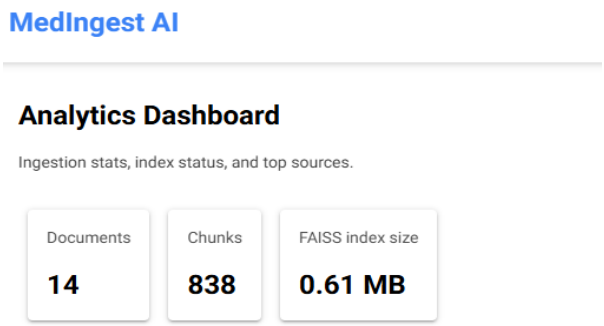


*Figure 6: System Analytics and Metrics View*

The analytics dashboard provides insights into system performance. It includes information such as number of processed documents, chunk distribution, and indexing details. This helps users understand how data is being handled internally and supports system monitoring.

**Export Functionality**

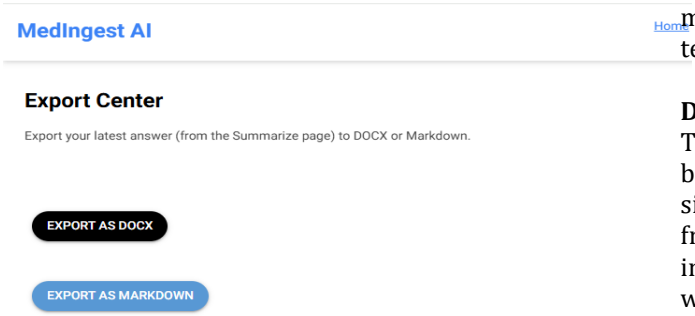


*Figure 7: Export Options Interface*

This interface allows users to export generated summaries into different formats such as DOCX or Markdown. This feature is useful for researchers who want to use the generated content in reports or academic work.

**System Workflow Visualization**

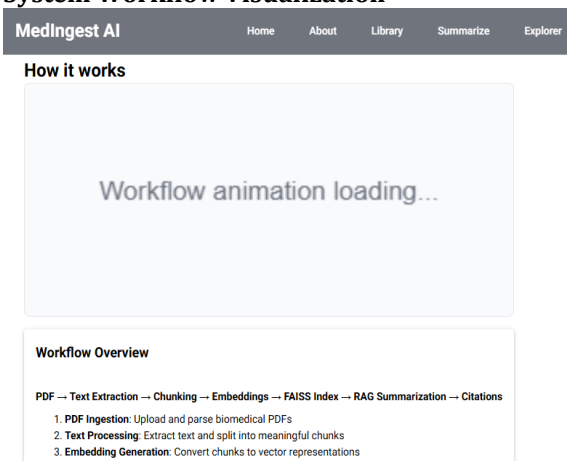


*Figure 8: End-to-End Workflow Execution Screen*

This screen represents the overall execution flow of the system, showing how data moves from document upload to final summary generation. It helps users understand the internal process and provides transparency in system operations.

**Implementation Summary**

The implementation demonstrates how different modules work together in a unified system. Each interface component is connected to backend processing stages such as ingestion, retrieval, and summarization. The use of visual dashboards and structured interfaces improves usability and makes the system accessible even to non-technical users.

**Discussion**

The developed system demonstrates how biomedical document processing can be simplified through an integrated and user-friendly platform. By combining document ingestion, semantic retrieval, and summarization within a single workflow, the system reduces the effort required to extract meaningful insights from large volumes of data. The presence of an interactive interface further improves usability, allowing users to perform complex operations without needing deep technical knowledge.

One of the key strengths of the system is its offline operation. Since all processing is performed locally, it eliminates concerns related to data privacy and dependency on external services. This makes the system suitable for environments where sensitive biomedical data is handled or where internet access is limited. The ability to run on standard hardware also increases accessibility for academic and research institutions.

The system also improves transparency by linking generated summaries with source content. Unlike generic AI-based tools, which often produce outputs without clear justification, this system maintains a connection between input documents and generated results. This helps users verify information and builds trust in the system's outputs.

Another important aspect is the modular design of the system. Each component—such as ingestion, indexing, retrieval, and summarization—works independently but contributes to the overall workflow. This modular structure makes it easier to maintain, extend, and upgrade the system in the future. For example, new retrieval methods or summarization techniques can be integrated without redesigning the entire system.

From a practical perspective, the inclusion of dashboards, document management, and export features makes the system highly usable in real-world scenarios. Researchers can manage large document collections, generate summaries, and directly use the output in their work. This reduces the gap between theoretical system design and actual implementation.

However, the system also has certain limitations. Since it operates offline, performance depends on local hardware capabilities. Additionally, the quality of summaries is influenced by the quality of uploaded documents and the effectiveness of the retrieval process. These factors highlight areas where further improvements can be explored in future work.

Overall, the system provides a balanced approach between functionality, usability, and reliability. It demonstrates that combining structured processing with an intuitive interface can significantly improve the efficiency of biomedical knowledge extraction.

## Conclusion

This work presents MedIngest AI, an offline biomedical knowledge engine designed to simplify document processing and generate evidence-supported summaries. The system integrates document ingestion, semantic retrieval, and summarization into a unified platform, making it easier for users to extract relevant information from large collections of biomedical data.

The implementation demonstrates that a structured workflow, combined with an interactive user interface, can significantly improve usability and efficiency. Features such as document management, query-based summarization, analytics dashboards, and export functionality make the system practical for real-world applications. The ability to link summaries with source content further enhances transparency and trust.

A key contribution of this system is its offline operation, which ensures data privacy and independence from external services. This makes it suitable for research environments where secure and reproducible processing is required. The modular design also allows flexibility for future enhancements and integration of additional features.

In summary, the system provides a reliable and user-friendly solution for biomedical knowledge extraction. It bridges the gap between complex document processing techniques and practical implementation, enabling users to work efficiently with large-scale biomedical information in a controlled and secure environment.

## References

Y.-H. Ke, L. Jin, K. Elangovan, H. R. Abdullah, N. Liu, and A. T. H. Sia, "Retrieval-augmented generation for 10 large language models and its generalizability in assessing medical fitness and preoperative instructions," npj Digital Medicine, 2025. DOI: 10.1038/s41746-025-01519-z.

L. Bednarczyk et al., "Scientific evidence for clinical text summarization using large language models: Scoping review," Journal of Medical Internet Research, vol. 27, 2025, Art. no. e68998. DOI: 10.2196/68998.

G. Zhang et al., "Leveraging long context in retrieval-augmented language models for medical applications," npj Digital Medicine, 2025. DOI: 10.1038/s41746-025-01651-w.

R. Yang et al., "Retrieval-augmented generation for generative artificial intelligence in health care: Equity, reliability, and personalization," npj Health Systems, 2025. DOI: 10.1038/s44401-024-00004-1.

A. Fink, D. A. Weinert, M. Bosma, and R. M. Summers, "Retrieval-Augmented Generation with Large Language Models: From theory to practice," Radiology: Artificial Intelligence, vol. 7, no. 4, 2025. DOI: 10.1148/ryai.240790.

D. A. Weinert et al., "Enhancing large language models with retrieval-augmented generation: A radiology-specific approach," Radiology: Artificial

Intelligence, vol. 7, no. 4, 2025. DOI: 10.1148/ryai.240313.

A. Wada et al., "Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation," npj Digital Medicine, vol. 8, 2025, Art. no. 395. DOI: 10.1038/s41746-025-01802-z.

M. Alkhalaf et al., "Applying generative AI with retrieval-augmented generation to summarize and extract key clinical information from electronic health records," Journal of Biomedical Informatics, vol. 156, 2024, Art. no. 104662. DOI: 10.1016/j.jbi.2024.104662.

M. Kirmani, A. Sinha, A. Bhattacharya, and D. Gupta, "Biomedical semantic text summarizer," BMC Bioinformatics, vol. 25, 2024, Art. no. 57. DOI: 10.1186/s12859-024-05712-x.

A. Givchi, R. Ramezani, and A. Baraani-Dastjerdi, "Graph-based abstractive biomedical text summarization," Journal of Biomedical Informatics, vol. 132, 2022, Art. no. 104099. DOI: 10.1016/j.jbi.2022.104099.

M. Wang, S. Luo, H. Xu, and Q. Hu, "A systematic review of automatic text summarization for biomedical literature and electronic health records," Journal of the American Medical Informatics Association, vol. 28, no. 10, pp. 2287–2297, 2021. DOI: 10.1093/jamia/ocab139.

A. Krithara et al., "BioASQ-QA: A manually curated corpus for biomedical question answering," Scientific Data, vol. 10, 2023, Art. no. 170. DOI: 10.1038/s41597-023-02068-4.

A. Nentidis et al., "Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF 2021," in CLEF 2021 Working Notes (LNCS/CLEF), 2021. [Online]. Available: https://ceur-ws.org/Vol-2936/paper-10.pdf.

M. Afzal, W. Wang, and H. Liu, "Clinical context–aware biomedical text summarization using PICO-based quality model," Journal of Medical Internet Research, vol. 22, no. 10, 2020, Art. no. e19810. DOI: 10.2196/19810.

M. Nasr Azadani, S. Minaei-Bidgoli, and H. Parvin, "Graph-based biomedical text summarization: An itemset mining approach," Journal of Biomedical Informatics, vol. 85, pp. 54–70, 2018. DOI: 10.1016/j.jbi.2018.06.005.

C. Hark, R. S. K. Singh, A. Rai, and U. Garain, "BioGraphSum: A graph-based model for biomedical text summarization," Heliyon, vol. 10, no. 10, 2024, e29509. DOI: 10.1016/j.heliyon.2024.e29509.

S. Liu et al., "Improving LLM applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines," Journal of the American Medical Informatics Association (JAMIA), 2025 (in press/online). [Online]. Available: https://amia.secure-platform.com/symposium/gallery/rounds/82021/details/19667

Z. Huang et al., "Biomedical automatic text summarization with large language models: A survey," Information Processing & Management, 2025. DOI: 10.1016/j.ipm.2025.103803.

D. Gupta, S. M. Ali, A. S. Chauhan, and S. Chakraborty, "A dataset of medical questions paired with automatically and manually verified answers and supporting scientific abstracts," Scientific Data, 2025. DOI: 10.1038/s41597-025-05233-z.

L. M. Amugongo, "Retrieval-augmented generation for large language models in healthcare: A review," PLOS Digital Health, 2025. DOI: 10.1371/journal.pdig.0000877.