ICRAMCS 2026

THE EIGHTH EDITION OF THE INTERNATIONAL CONFERENCE ON RESEARCH IN APPLIED MATHEMATICS AND COMPUTER SCIENCE
April 23-24-25, 2026 | Marrakech, Morocco

Retrieval-Augmented Generation in Biomedicine: A Survey of Methods and Architectures

Communication Info

Authors:
Sara BARKAOUI[1]
Sara BOUZID[1]
Abdelghafour ATLAS [1]

[1] Cadi Ayyad University, UCA, ENSA Marrakech, LMSC Laboratory, Marrakech, Morocco.

Keywords:
(1) LLM
(2) RAG
(3) Biomedical QA

Abstract

Large language models have recently demonstrated good performance in various tasks related to automatic natural language understanding and generation. However, their use in medical and clinical situations raises critical concerns regarding hallucinations, lack of information tracking, and outdated knowledge [1,2]. The retrieval-augmented generation (RAG) method aims to solve these challenges by combining external information with LLM capabilities to produce specialized and up-to-date responses [3].

This survey provides a structured overview of RAG-based methods in the biomedical field. By carefully analyzing eighteen major studies, we developed a classification framework that categorizes these systems according to their knowledge bases, information retrieval mechanisms, linguistic model integration architectures, and application domains. A comparative evaluation highlights current trends in the field, limitations of existing solutions, and unresolved questions regarding the implementation of robust RAG systems in the biomedical context. This study shows that in biomedical applications, RAG significantly improves response accuracy and domain relevance. However, performance critically depends on effective context construction using techniques like segmentation, classification, and semantic clustering, which directly influence the consistency and reliability of generated outputs [4,5].

References

[1] Öğdü ÇU, Arslanoğlu K, Karaköse M. An Adaptive Multi-Agent LLM-Based Clinical Decision Support System Integrating Biomedical RAG and Web Intelligence. *IEEE Access*, 13, 2025, pp. 167390–404.

[2] Soman K, Rose PW, Morris JH, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40, 2024, btae560.

[3] Zhan Z, Wang J, Zhou S, et al. MMRAG: Multi-Mode Retrieval-Augmented Generation with Large Language Models for Biomedical In-Context Learning. *Journal of the American Medical Informatics Association*, 2025.

[4] Gao Y, Zong L, Li Y. *Enhancing Biomedical Question Answering with Parameter-Efficient Fine-Tuning and Hierarchical Retrieval Augmented Generation.* In: CLEF 2024 – 12th BioASQ Challenge Workshop. CEUR Workshop Proceedings, 2024.

[5] Matsumoto N, Choi H, Moran J, et al. *ESCARGOT: an AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning.* Bioinformatics, 41, 2025, btaf031.