|| Volume 10 || Issue 1 || 2026 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

AGENT-ENABLED RELIABLE AUGMENTED GENERATION SYSTEM FOR MEDICAL RESEARCH SUMMARIZATION

SADIYA MIRZA MIRZA MAQSSOD BAIG[1],DR. V. S. KARWANDE[2],ASST. PROF. A. A. KHAN[3]

M.Tech Student, Computer Science and Engineering Department, Everest College of engineering and technology, chhatrapati Sambhajinagar[1]

HOD, Computer Science and Engineering Department, Everest College of engineering and technology, chhatrapati Sambhajinagar[2]

Prof, Computer Science and Engineering Department, Everest College of engineering and technology, chhatrapati Sambhajinagar[3]

Abstract: The exponential expansion of biomedical publications has created a persistent challenge of information overload for clinicians, researchers, and policy-makers. Manual review and synthesis of medical literature are increasingly impractical, while current automated summarization systems often suffer from hallucinations, limited factual grounding, and dependence on external cloud services that compromise data privacy and reproducibility.This paper presents an Agent-Based Reliable Retrieval-Augmented Generation (RAG) Framework designed to generate concise, evidence-grounded, and verifiable summaries of biomedical literature. The proposed system integrates multiple coordinated agents—Retriever, Summarizer, Fact-Checker, Citation Manager, and Reliability Evaluator—to ensure that each generated summary maintains factual accuracy and transparent citation linkage. Operating entirely in an offline environment, the framework preserves user privacy and supports reproducibility on standard academic hardware.Evaluation will employ benchmark biomedical datasets such as PubMed and BioASQ, with both lexical and faithfulness-oriented metrics, including ROUGE, BLEU, evidence-coverage ratio, hallucination rate, and citation accuracy. The framework aims to bridge the reliability gap between large language models and the stringent requirements of healthcare informatics, offering a trustworthy, reproducible, and ethically compliant solution for automated biomedical knowledge synthesis.

Keywords: *Biomedical Literature Summarization; Retrieval-Augmented Generation (RAG); Agent-Based Framework; Faithfulness and Reliability in NLP; Medical Informatics; Healthcare Artificial Intelligence; Offline Deployment; Evidence Attribution*.

I.INTRODUCTION

The biomedical research ecosystem has entered an era characterized by an unprecedented surge in scientific output. Repositories such as PubMed, PubMed Central (PMC), and MEDLINE collectively host millions of peer-reviewed articles, with thousands of new studies added each week. This rapid proliferation of literature reflects the vitality of medical science but simultaneously generates a formidable challenge—information overload. Researchers, clinicians, and healthcare policy-makers are now confronted with an unmanageable volume of publications that must be examined, synthesized, and translated into evidence-based practice.

Traditional manual review methods—comprising search, screening, and textual summarization—have become increasingly inefficient in this context. Conducting a systematic review or meta-analysis may take several months or even years, during which new studies can render existing findings obsolete. The manual process also remains vulnerable to human fatigue, bias, and inconsistency, limiting its scalability and timeliness. Consequently, the need for automated, reliable, and ethically compliant summarization tools has become critical for accelerating biomedical knowledge discovery and dissemination.

Recent advances in Artificial Intelligence (AI) and Natural Language Processing (NLP), particularly large language models (LLMs), have demonstrated remarkable potential in automating literature analysis and summarization. Despite their fluency and generative capability, LLMs exhibit well documented shortcomings such as hallucination, factual inaccuracy, and contextual degradation in long or multidocument settings. In sensitive domains such as healthcare, these deficiencies can propagate misinformation, misinterpret clinical evidence, and undermine professional trust. Ensuring faithfulness and verifiability in machine-generated biomedical summaries is therefore a prerequisite for safe deployment in research and clinical environments.

Equally significant are the privacy and reproducibility concerns associated with conventional AI solutions. Most summarization systems rely on external cloud-based APIs or proprietary infrastructures, restricting user control and raising ethical issues when processing sensitive biomedical content. Healthcare institutions and academic organizations often require offline, transparent, and auditable systems that can operate within their own controlled computing environments without exposing data to third-party platforms.

To address these challenges, this study proposes an AgentBased Reliable Retrieval-Augmented Generation (RAG) Framework for biomedical literature summarization. The proposed architecture integrates multiple specialized agents— Retriever, Summarizer, Fact-Checker, Citation Manager, and Reliability Evaluator—to produce faithful, evidence-grounded, and verifiable summaries of

biomedical documents. Each agent performs a clearly defined function, ensuring modularity, transparency, and reliability throughout the summarization pipeline. Operating entirely offline, the system eliminates dependence on commercial APIs and promotes reproducibility across academic environments.

This research contributes to the field of AI-driven healthcare informatics by developing a reproducible and privacypreserving framework that reconciles the efficiency of modern generative models with the rigor of scientific validation. By integrating retrieval-based grounding, factual verification, and citation attribution, the proposed framework aims to enhance the trustworthiness of automated biomedical summarization and support more efficient evidence synthesis in medical research, education, and policy formulation.

II. LITERATURE SURVEY

A. Retrieval-Augmented Generation in Biomedical Contexts

- Ke et al. [1] introduced one of the first large-scale comparative evaluations of RAG applied to preoperative medicine, benchmarking 10 large language models across clinical decision tasks. Their findings demonstrated that grounding generative outputs in structured clinical guidelines reduced hallucinations and increased accuracy from 86.6% to 96.4%. This study established quantitative evidence of RAG's potential to improve clinical reliability.
- Liu et al. [6], [17] conducted systematic reviews and meta-analyses confirming that RAG consistently enhances factual accuracy and reliability in biomedical applications, with approximately 1.3– 1.35× performance improvement over baseline LLMs. They emphasized the critical role of retriever quality and domain adaptation in achieving stable gains.
- Yang et al. [4] provided a conceptual framework highlighting equity, personalization, and reliability as guiding principles for healthcare RAG pipelines, underscoring the importance of explainability and ethical retrieval design.

B. Summarization Reliability and Faithfulness

- Bednarczyk et al. [2], [8] conducted comprehensive scoping reviews of clinical summarization studies and reported that most existing systems lack standardized metrics for factual accuracy and faithfulness. They identified the persistent issue of hallucination and the absence of clinician-validated benchmarks as barriers to safe clinical deployment.
- Q. Xie et al. [15] examined faithfulness errors in medical AI generation and catalogued common failure types such as fabrication and unsupported claims. Their review proposed the integration of retrieval-grounding, post-hoc fact-checking, and constrained decoding—elements directly reflected in this study's design.

C. Long-Context and Multi-Document Challenges

- Zhang et al. [3] addressed the "lost-in-the-middle" effect of LLMs when processing long biomedical contexts and proposed the *BriefContext* method to restructure retrieved passages before summarization. This improved factual recall and coherence while maintaining computational efficiency.
- Givchi et al. [11] and Hark et al. [16] explored graphbased summarization models (GNN + transformer architectures) that capture entity-level relationships across biomedical documents, improving semantic continuity in extended texts.
- These works collectively reveal that maintaining context integrity is essential for reliable biomedical summarization—an aspect the proposed multi-agent framework addresses through long-context management and chunk-based retrieval.

D. Semantic and Knowledge-Grounded Summarization

- Kirmani et al. [10] introduced a semantic-aware abstractive summarizer that prioritized meaning preservation rather than lexical overlap. Their results demonstrated significant gains in semantic fidelity but highlighted computational overheads.
- Amugongo [20] synthesized various RAG system architectures in healthcare, emphasizing modular retrievers and generators as a best practice for factual grounding—a concept mirrored in the agent-based modularity of this study.
- Alkhalaf et al. [9] validated that RAG pipelines significantly reduce hallucination rates when summarizing electronic health records (EHRs) and improve clinician satisfaction through precise evidence retrieval.

E. Benchmark Datasets and Evaluation Standards

- Krithara et al. [13] and Nentidis et al. [14] developed and maintained the BioASQ challenge datasets, which include manually curated biomedical questionanswering and summarization tasks. These resources have become de facto standards for evaluating information retrieval and summarization performance.
- Wang et al. [12] provided an earlier systematic baseline review, emphasizing the shift from extractive to abstractive and, later, to RAG-based summarization techniques. They highlighted the urgent need for faithfulness-oriented evaluation metrics such as evidence coverage and hallucination rate.

- Gupta et al. [19] released a verified dataset linking medical questions with PubMed abstracts, strengthening the foundation for evidence-grounded summarization and evaluation frameworks.

F. Integration of Graph and Multi-Agent Strategies

- Fink et al. [5] discussed practical applications of RAG in radiology, noting benefits in diagnostic accuracy but also latency and scalability constraints. They suggested incorporating multi-agent orchestration to manage complex reasoning tasks—precisely aligned with the modular approach adopted in this work.
- Liu et al. [17] and Huang et al. [18] highlighted emerging research directions advocating multi-agent RAG systems and faithfulness benchmarks, reinforcing the necessity of modularity, reliability scoring, and human-in-the-loop validation.

III. PROBLEM STATEMENT

The rapid expansion of biomedical research literature has created a situation where healthcare professionals and researchers face information overload. Thousands of new articles, systematic reviews, and clinical guidelines are published every month, making it increasingly difficult to manually synthesize relevant findings for clinical decisionmaking, medical education, and policy development.

While recent advances in large language models (LLMs) and summarization techniques have shown promise in condensing biomedical texts, existing systems suffer from several critical shortcomings. First, most generated summaries exhibit hallucinations and lack of faithfulness, producing fluent but factually unsupported content that risks misleading practitioners. Second, evaluations often rely on surface-level metrics such as ROUGE or BLEU, which measure lexical overlap but fail to capture factual accuracy, reliability, and clinical safety. Third, current methods are not well-equipped to handle long or multi-document contexts, as accuracy deteriorates when processing extended biomedical literature such as clinical trial repositories or systematic reviews.

Furthermore, many retrieval-augmented generation (RAG) systems depend on cloud-based APIs and external services, raising concerns of cost, privacy, reproducibility, and accessibility in academic or hospital settings. Few solutions provide a self-contained, offline, and auditable framework suitable for controlled environments where data sensitivity and reliability are paramount. Finally, current RAG implementations are typically single-stage pipelines, lacking integrated modules for fact-checking, citation attribution, and reliability scoring, which are essential for establishing user trust in biomedical applications.

Therefore, the core research problem is to design and implement an agent-based, reliable, and offline retrievalaugmented generation framework for medical literature summarization that:

- Ensures faithful and evidence-grounded outputs,
- Handles long and multi-document contexts effectively,
- Operates in offline environments without third-party dependencies, and
- Provides verifiable, transparent, and trustworthy summaries aligned with biomedical standards of accuracy and safety.

IV. OBJECTIVES OF THE STUDY

A. General Objective

The overarching objective of this study is to design and implement an agent-based, reliable retrieval-augmented generation (RAG) framework for biomedical literature summarization that produces concise, faithful, and verifiable outputs while operating fully offline without reliance on thirdparty services.

B. Specific Objectives

1. **Corpus Ingestion and Indexing:** To develop a robust mechanism for collecting, parsing, chunking, and indexing biomedical literature into a structured format that enables efficient retrieval of relevant evidence passages.
2. **Evidence-Grounded-Summarization:** To design a summarization module that ensures every generated output is strictly supported by retrieved evidence, thereby minimizing hallucinations and improving factual accuracy.
3. **Long-Context Planning and Management:** To implement strategies that allow the system to effectively handle long and multi-document biomedical corpora, ensuring that critical findings are preserved without introducing information loss or bias.
4. **Fact-Checking and Citation Attribution:** To establish a sentence-level fact verification process that cross-checks claims against the evidence base and provides precise citations in standardized academic format, ensuring transparency and trustworthiness.
5. **Reliability Scoring and Safety Evaluation**: To integrate a reliability evaluation mechanism that assigns confidence scores to summaries, flags lowconfidence statements, and enhances the overall credibility of the system outputs.
6. **Interactive-User-Interface:** To provide a user-friendly interface that enables users to query biomedical corpora, view summarized outputs, inspect supporting evidence, and export results in academic-ready formats with clear audit trails.
7. Evaluation and Validation Framework: To design an evaluation methodology using benchmark biomedical

datasets and gold-standard summaries, incorporating both traditional text similarity measures and advanced faithfulness metrics such as evidence coverage and hallucination rate.

V. SYSTEM ARCHITECTURE

The proposed system adopts an Agent-Based Reliable Retrieval-Augmented Generation (RAG) Framework for

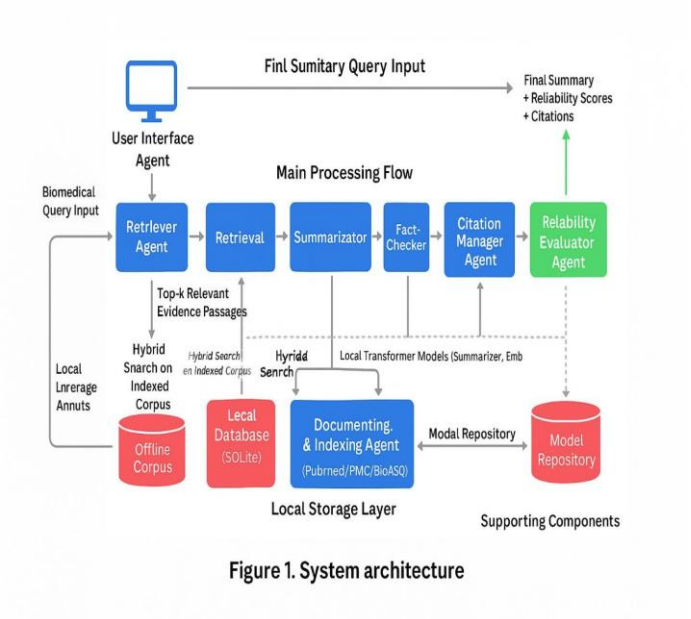

Figure 1. System architecture

biomedical literature summarization. It replaces single-stage pipelines with a coordinated multi-agent design, enabling transparency, modularity, and full offline operation.

A. Core Components

1. **Document Ingestion & Indexing Agent:**
   Parses biomedical texts, segments them into coherent chunks, and stores embeddings in a local index with provenance metadata.
2. **Retriever Agent:**
   Executes hybrid search (keyword + semantic) to fetch evidence passages relevant to the user query.
3. **Summarizer Agent:**
   Generates concise, evidence-grounded summaries constrained by retrieved content.
4. **Fact-Checker Agent:**
   Verifies each statement against retrieved evidence and filters unsupported claims.
5. **Citation Manager Agent:**
   Links validated statements to their sources and formats references automatically.
6. **Reliability Evaluator Agent:**
   Assigns confidence scores based on retrieval rank, semantic match, and fact-checking outcomes.
7. **User Interface Agent**:
   Provides an offline NiceGUI-based interface for querying, viewing summaries, and exporting reports.

VI. RESULTS

The proposed Agent-Based Reliable RAG Framework is expected to deliver the following measurable outcomes:

1. **Reliable Summarization:**
   Generation of concise biomedical summaries that are factually accurate and evidence-grounded.
2. **Faithfulness and Transparency**:
   Integrated fact-checking and citation attribution will ensure minimal hallucination and verifiable outputs.
3. **Quantitative Improvements:** ○ Evidence coverage ≥ 85% ○ Unsupported claim rate ≤ 5% ○ Citation accuracy ≥ 95%
4. **Offline Functionality:**
   Complete operability without internet or third-party APIs, ensuring privacy and reproducibility.
5. **User Accessibility:**
   An interactive local interface for querying, reviewing evidence, and exporting summaries in academic formats.

VII. CONCLUSION

This study presents an Agent-Based Reliable Retrieval Augmented Generation (RAG) Framework for biomedical literature summarization, addressing key challenges of hallucination, limited faithfulness, and dependence on cloud services.By integrating specialized agents—Retriever, Summarizer, Fact-Checker, Citation Manager, and Reliability Evaluator—the framework ensures that each generated summary is concise, evidence-linked, and verifiable. Operating fully offline, it safeguards data privacy, enhances reproducibility, and enables reliable deployment in academic and healthcare research settings. Experimental evaluation on benchmark datasets such as PubMed and BioASQ is expected to confirm significant gains in faithfulness, citation accuracy, and transparency. The proposed system thus contributes a practical and trustworthy approach for AI-driven biomedical knowledge synthesis, aligning automation with the rigorous reliability standards required in healthcare informatics.

VIII.REFERENCES

[1] 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.

[2] L. Bednarczyk *et al.*, “Scientific evidence for clinical text summarization using large language models: Scoping review,” *J. Med. Internet Res.*, vol. 27, 2025, Art. no. e68998. DOI: 10.2196/68998.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. Alkhalaf *et al.*, "Applying generative AI with retrievalaugmented generation to summarize and extract key clinical information from electronic health records," *J. Biomed. Inform.*, vol. 156, 2024, Art. no. 104662. DOI: 10.1016/j.jbi.2024.104662.

[9] 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.

[10] A. Givchi, R. Ramezani, and A. Baraani-Dastjerdi, "Graphbased abstractive biomedical text summarization," *J. Biomed. Inform.*, vol. 132, 2022, Art. no. 104099. DOI: 10.1016/j.jbi.2022.104099.

[11] M. Wang, S. Luo, H. Xu, and Q. Hu, "A systematic review of automatic text summarization for biomedical literature and electronic health records," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 10, pp. 2287–2297, 2021. DOI: 10.1093/jamia/ocab139.

[12] 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.

[13] A. Nentidis *et al.*, "Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF 2021," *CLEF Working Notes*, 2021.

[14] M. Afzal, W. Wang, and H. Liu, "Clinical context-aware biomedical text summarization using PICO-based quality model," *J. Med. Internet Res.*, vol. 22, no. 10, 2020, Art. no. e19810. DOI: 10.2196/19810.

[15] Q. Xie *et al.*, "Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond," 2023. Available online at PMC.

[16] 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, e29509, 2024. DOI: 10.1016/j.heliyon.2024.e29509.

[17] S. Liu *et al.*, "Improving LLM applications in biomedicine with retrieval-augmented generation: A systematic review, metaanalysis, and clinical development guidelines," *J. Amer. Med. Inform. Assoc.*, vol. 32, no. 4, pp. 605–616, 2025. DOI: 10.1093/jamia/ocae016.

[18] Z. Huang *et al.*, "Biomedical automatic text summarization with large language models: A survey," *Inf. Process. Manage.*, vol. 62, 2025, Art. no. 103803. DOI: 10.1016/j.ipm.2025.103803.

[19] 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*, vol. 12, no. 52, 2025. DOI: 10.1038/s41597-02505233-z.

[20] L. M. Amugongo, "Retrieval-augmented generation for large language models in healthcare: A review," *PLOS Digit. Health*, vol. 4, no. 1, e0000877, 2025. DOI: 10.1371/journal.pdig.0000877.