

A Review of Metaheuristic Optimization Techniques for Improving Retrieval Accuracy and Response Reliability in Healthcare RAG Systems

P.G.Muraleedharan¹ and Dr. Vikas Nana Mahandule²

¹Research Scholar, Department of Computer Science & IT

²Research Guide, Department of Computer Science & IT

Sunrise University, Alwar, Raj., India

Abstract: Retrieval-Augmented Generation (RAG) has emerged as a promising artificial intelligence framework for integrating large language models (LLMs) with external knowledge sources to generate contextually relevant, evidence-supported responses. In healthcare, RAG systems can retrieve information from clinical guidelines, biomedical literature, electronic health records, drug databases, and medical knowledge repositories before generating answers to clinical questions. However, their effectiveness depends on the quality of document retrieval, relevance of retrieved evidence, completeness of contextual information, and reliability of the generated response. Conventional retrieval approaches may struggle with complex medical terminology, ambiguous queries, outdated information, conflicting evidence, and domain-specific knowledge requirements..

Keywords: Retrieval-Augmented Generation, Information Retrieval, Response Reliability, Large Language Models

I. INTRODUCTION

Artificial intelligence has become increasingly important in healthcare because of its potential to support clinical information retrieval, medical research, patient education, diagnostic reasoning, and clinical decision support. Large language models have demonstrated substantial capabilities in understanding natural language and generating coherent responses. Nevertheless, their practical use in medicine presents challenges related to factual accuracy, outdated knowledge, unsupported conclusions, and the generation of plausible but incorrect information. These limitations are particularly important in clinical environments, where incorrect recommendations can affect patient safety and treatment decisions.

Retrieval-Augmented Generation provides a mechanism for reducing some of these limitations by connecting a language model to external information sources. Instead of generating a response exclusively from its learned parameters, a RAG system retrieves relevant documents from a knowledge base and incorporates the retrieved information into the generation process. Lewis et al. (2020) introduced a foundational RAG framework that combines parametric language-model knowledge with non-parametric retrieved knowledge. This approach established an important direction for developing systems that can generate answers using information obtained from external documents.

In healthcare, RAG can support medical question answering, guideline summarization, pharmaceutical information retrieval, evidence-based clinical assistance, and biomedical literature analysis. For example, a system responding to a question about medication contraindications could retrieve relevant prescribing information and clinical guidelines before generating a concise explanation. However, the system's reliability depends on whether the retrieval component identifies the correct documents and whether the language model interprets the evidence appropriately. If the retrieved

information is incomplete, irrelevant, outdated, or contradictory, the final response may remain unreliable despite its fluent presentation.

Metaheuristic optimization offers a promising approach to addressing these problems. Metaheuristics are general-purpose search strategies designed to identify high-quality solutions in complex optimization spaces. Unlike many conventional optimization methods, they can operate without requiring gradient information and can accommodate nonlinear, discontinuous, or competing objective functions. Genetic algorithms explore candidate solutions through selection, crossover, and mutation, whereas particle swarm optimization uses the collective search behavior of particles to improve candidate solutions. Other methods, including ant colony optimization, differential evolution, and simulated annealing, provide alternative mechanisms for exploring parameter configurations.

Within healthcare RAG systems, metaheuristics may optimize document chunk size, chunk overlap, embedding configurations, retrieval depth, hybrid-search weights, reranking thresholds, query-rewriting parameters, and evidence-selection strategies. They may also help balance retrieval quality against computational cost and response latency. Importantly, these applications represent optimization opportunities rather than guaranteed improvements: their effectiveness must be established experimentally against appropriate baselines.

The central argument of this review is that healthcare RAG reliability requires more than improving retrieval scores alone. A useful system must retrieve relevant evidence, preserve clinically important details, identify uncertainty, avoid unsupported recommendations, and communicate limitations clearly. Therefore, metaheuristic optimization should be evaluated through a multi-objective framework that considers retrieval effectiveness, factual grounding, clinical completeness, safety, and computational efficiency. Such an approach can provide a more comprehensive foundation for developing dependable healthcare information systems.

METAHEURISTIC OPTIMIZATION TECHNIQUES

Metaheuristic algorithms search for high-quality solutions in complex parameter spaces. In a healthcare RAG environment, a candidate solution may represent a particular combination of chunk size, embedding model, retrieval depth, search weights, reranking threshold, and generation settings. A fitness function evaluates each configuration according to predefined performance objectives.

1. Genetic Algorithm

A genetic algorithm (GA) imitates evolutionary processes through selection, crossover, and mutation. Initially, multiple candidate retrieval configurations are generated. Each configuration is evaluated using a fitness function based on retrieval relevance, evidence coverage, response quality, and computational cost. Better-performing candidates are selected to produce subsequent generations.

In healthcare RAG, GA may be used to select combinations of retrieval parameters or identify useful feature subsets for medical document ranking. Its population-based search can explore multiple configurations simultaneously, although the evaluation of many candidates may require substantial computational resources.

2. Particle Swarm Optimization

Particle swarm optimization (PSO) uses a population of candidate solutions, called particles, that adjust their positions based on their own best-known solutions and the best solutions identified by the swarm. PSO may be suitable for tuning continuous parameters, including hybrid-retrieval weights, similarity thresholds, and weighting coefficients in ranking functions.

For example, a system combining dense semantic retrieval with keyword-based retrieval can optimize the relative contribution of each retrieval score. The resulting weights should be evaluated on independent medical queries to ensure that optimization improves generalization rather than merely fitting the development dataset.

3. Ant Colony Optimization

Ant colony optimization (ACO) is inspired by the collective foraging behavior of ants. It uses accumulated search information to guide the selection of promising paths through a problem space. In RAG systems, ACO could be

investigated for selecting sequences of retrieval operations, navigating knowledge graphs, or identifying combinations of medical evidence sources.

For example, a complex clinical question may require evidence from a treatment guideline, a drug interaction database, and a disease-specific reference. An ACO-based approach could explore alternative evidence-selection pathways. However, its practical value depends on the quality of the graph representation, the design of the search space, and the computational overhead.

4. Differential Evolution

Differential evolution (DE) is a population-based optimization method particularly suited to continuous parameter optimization. Candidate solutions are modified using differences between population members, followed by selection based on objective-function performance.

For healthcare RAG, DE may optimize retrieval weights, ranking thresholds, and combinations of performance parameters. It is especially relevant when the fitness function is nonlinear or when gradients are unavailable. Nevertheless, discrete design decisions, such as selecting a particular embedding model, require appropriate encoding or a mixed-variable optimization strategy.

5. Simulated Annealing

Simulated annealing (SA) searches for improved solutions by exploring neighboring configurations. It sometimes accepts a worse solution to escape local optima, with the likelihood of doing so decreasing as the search progresses.

SA could optimize retrieval depth, chunking parameters, or query-processing configurations when the search space is relatively manageable. Its implementation is comparatively straightforward, but its effectiveness depends on the neighborhood definition, cooling schedule, and number of evaluations.

6. Multi-Objective and Hybrid Optimization

Healthcare RAG cannot be optimized effectively through a single performance measure. A system with excellent retrieval recall may still be slow, return excessive irrelevant material, or generate responses that are insufficiently supported by evidence. Multi-objective optimization addresses these competing requirements by evaluating several objectives together.

A hybrid method could combine GA for discrete configuration selection with PSO or DE for continuous parameter tuning. Another approach could combine metaheuristic search with Bayesian optimization to allocate expensive evaluations more efficiently. These combinations are proposed research directions and should not be interpreted as established clinical performance improvements without comparative testing.

IMPROVING RETRIEVAL ACCURACY THROUGH METAHEURISTIC OPTIMIZATION

Retrieval accuracy is a fundamental determinant of RAG performance because the generation model can only use the evidence made available to it during retrieval. Even a powerful language model may generate an incomplete or incorrect response when the retrieval stage fails to identify the relevant medical information.

1. Optimization of Document Chunking

Document chunking divides long medical documents into smaller units that can be retrieved independently. The effectiveness of chunking depends on document structure, terminology, clinical context, and the nature of the query. A treatment guideline, for example, may contain a recommendation whose meaning depends on a nearby contraindication or qualifying statement.

A metaheuristic algorithm could search for suitable chunk sizes and overlap values by evaluating retrieval relevance and contextual completeness across a development dataset. The fitness function should penalize configurations that retrieve relevant keywords but omit essential clinical qualifications. The optimized settings should then be tested on unseen documents and queries.

2. Optimization of Hybrid Retrieval

Healthcare terminology often includes abbreviations, synonyms, drug names, disease classifications, and specialized clinical expressions. Dense retrieval can capture semantic similarity, whereas keyword-based retrieval can be useful

when exact terminology matters. Combining the two approaches may improve coverage, but the appropriate balance depends on the clinical domain.

A hybrid-ranking score can be expressed as:

$$S(d, q) = \alpha S_{\text{dense}}(d, q) + (1 - \alpha) S_{\text{keyword}}(d, q)$$

Here, $S(d, q)$ represents the combined relevance score of document (d) for query (q), and is a weight between zero and one. A metaheuristic algorithm can optimize this weight alongside retrieval depth and reranking thresholds. The objective is not simply to retrieve more documents, but to retrieve the most relevant and clinically informative evidence.

3. Optimization of Document Reranking

Initial retrieval may return several potentially relevant documents in an imperfect order. Reranking evaluates these candidates more carefully and prioritizes those that best address the question. Optimization may focus on the number of candidates retained, the balance between semantic and lexical scores, and the threshold used to exclude irrelevant evidence.

Clinical retrieval also requires attention to source authority, publication date, population characteristics, and guideline version. A recent but low-quality source should not automatically outrank an authoritative clinical guideline. Therefore, retrieval optimization should incorporate source-quality constraints instead of relying exclusively on similarity scores.

4. Query Reformulation and Expansion

Medical questions may contain misspellings, abbreviations, ambiguous expressions, or descriptions that differ from the language used in clinical documents. Query expansion can add relevant synonyms, terminology variants, or clinically meaningful concepts. However, excessive expansion can introduce unrelated information or change the original meaning.

Metaheuristic search could evaluate alternative query transformations according to retrieval coverage, relevance, and preservation of clinical intent. Any optimized reformulation method should be tested on both common and uncommon conditions, including queries containing negation, dosage constraints, contraindications, and population-specific requirements.

IMPROVING RESPONSE RELIABILITY AND EVIDENCE GROUNDING

Retrieval accuracy alone does not guarantee a reliable final answer. The generation model may misinterpret evidence, overlook exceptions, combine contradictory findings, or produce statements that are not supported by the retrieved documents.

Response reliability should therefore be assessed through several complementary dimensions:

Factual consistency: whether the generated claims agree with the retrieved evidence.

Citation correctness: whether each citation genuinely supports the associated claim.

Evidence completeness: whether clinically important information has been included.

Uncertainty handling: whether the system identifies insufficient or conflicting evidence.

Safety: whether the response avoids unsupported diagnosis or treatment recommendations.

Temporal validity: whether the information remains appropriate for the relevant date, guideline version, and clinical context.

A possible multi-objective optimization function is:

$$\max F(x) = (R(x), G(x), C(x))$$

subject to:

$$L(x) \leq L_{\max}, \quad K(x) \leq K_{\max}$$

where (x) represents a RAG configuration, (R) represents retrieval quality, (G) represents evidence grounding, (C) represents answer completeness, (L) represents latency, and (K) represents computational cost. The limits (K) represent application-specific constraints. Safety requirements should be treated as explicit acceptance criteria, not merely as objectives that can be traded away for speed or accuracy.

A multi-objective algorithm may produce a Pareto frontier, representing configurations for which improving one objective requires sacrificing another. Healthcare developers can then select a configuration that meets predefined safety and performance thresholds while maintaining acceptable computational efficiency.

Recent healthcare RAG reviews emphasize the need for stronger and more standardized evaluation of retrieved evidence, generated responses, and ethical considerations. These findings reinforce the importance of measuring retrieval and generation separately rather than assuming that improvements in one automatically improve the other.

EVALUATION METRICS FOR HEALTHCARE RAG SYSTEMS

An effective evaluation framework should measure retrieval performance, answer quality, factual grounding, clinical safety, and operational efficiency. Table 2 summarizes metrics suitable for a proposed experimental investigation.

Table 1. Evaluation metrics for metaheuristic-optimized healthcare RAG systems

Evaluation dimension	Metric or method	Purpose
Retrieval relevance	Precision@k	Measures the proportion of retrieved top-(k) documents that are relevant
Retrieval coverage	Recall@k	Measures the proportion of relevant documents retrieved within the top (k)
Ranking quality	Mean Reciprocal Rank (MRR)	Evaluates how highly the first relevant result is ranked
Ranking quality	Normalized Discounted Cumulative Gain (NDCG)	Evaluates ranking quality when relevance has multiple grades
Answer correctness	Expert-validated accuracy	Measures whether the answer correctly addresses the clinical question
Evidence grounding	Claim-level support rate	Measures the proportion of answer claims supported by cited evidence
Citation reliability	Citation precision and correctness	Checks whether citations point to sources that substantiate the claims
Completeness	Expert-rated evidence coverage	Assesses whether important clinical facts and qualifications are included
Safety	Clinical error and unsafe-recommendation rate	Identifies clinically consequential errors or unsafe outputs
Uncertainty	Abstention quality	Evaluates whether the system appropriately declines to answer when evidence is inadequate
Efficiency	Median and 95th-percentile latency	Measures typical and high-end response time
Resource use	Computational cost per query	Estimates the expense of retrieval, optimization and generation

Source note: The metrics are established evaluation concepts assembled here into a proposed healthcare RAG framework; the table does not report results from a newly conducted experiment.

Precision@k and Recall@k evaluate different aspects of retrieval performance. A system may achieve high recall by retrieving many documents but have poor precision because many results are irrelevant. MRR and NDCG help determine whether the most useful evidence appears near the top of the ranking.

Answer-level evaluation should be supplemented with claim-level verification. A response can be broadly correct while containing one unsupported statement about a medication contraindication or treatment threshold. Independent clinical

reviewers should therefore assess the correctness and relevance of individual claims, the adequacy of their supporting sources, and the potential consequences of errors.

The 2024 systematic review by Kell et al. on point-of-care question-answering systems also illustrates the importance of evaluating the practical performance of information systems intended for healthcare professionals.

PROPOSED RESEARCH FRAMEWORK

A rigorous investigation of metaheuristic optimization in healthcare RAG can be organized into the following stages.

Dataset development: Assemble representative medical questions and a curated collection of clinical guidelines, biomedical publications, and approved reference documents. Establish expert-validated relevance judgments and reference answers.

Baseline construction: Develop a conventional RAG system using a fixed embedding model, standard chunking, dense or hybrid retrieval, and a consistent generation model.

Optimization design: Implement selected algorithms, such as GA, PSO, DE, and SA, with clearly defined candidate parameters, search bounds, stopping criteria, and fitness functions.

Comparative testing: Compare each optimized system with the baseline and with conventional hyperparameter tuning. Use the same data partitions, retrieval corpus, generation model, and evaluation procedures wherever possible.

Reliability evaluation: Measure retrieval relevance, answer correctness, claim-level grounding, citation validity, completeness, and clinical safety using automated metrics and independent expert assessment.

Statistical analysis: Report confidence intervals, variability across repeated runs, and suitable paired statistical tests. Account for the computational budget used by each optimizer.

Robustness and generalization: Test the best configurations on unseen queries, different clinical specialties, updated documents, ambiguous questions, and conflicting evidence.

Clinical implementation assessment: Examine privacy, latency, interpretability, source traceability, monitoring requirements, and the need for human review before considering deployment.

For experimental fairness, optimization and evaluation datasets should be separated. Otherwise, a metaheuristic algorithm may overfit the fitness function and appear more effective than it would be on unseen clinical questions. Each optimizer should receive a comparable evaluation budget, and the baseline should be tuned fairly rather than deliberately left underoptimized.

The experimental analysis should also distinguish between retrieval improvement and downstream response improvement. For example, a statistically meaningful increase in Recall@5 does not necessarily imply better clinical accuracy if the newly retrieved documents are contradictory or difficult for the generation model to interpret.

DISCUSSION

The reviewed literature establishes a strong rationale for improving retrieval-augmented generation in healthcare, but the specific contribution of metaheuristic optimization remains an empirical research question. The healthcare RAG literature documents applications across medical question answering, clinical decision support, and biomedical research. However, differences in datasets, retrieval configurations, language models, and evaluation protocols make direct comparisons difficult. Amugongo et al. (2025), for example, identify substantial variation in healthcare RAG approaches and emphasize the need for more consistent evaluation and attention to ethical considerations.

Metaheuristic algorithms offer a potentially useful way to search for retrieval configurations that balance several competing objectives. Nevertheless, algorithm selection should be guided by the structure of the optimization problem, computational constraints, and measured performance. A more complicated optimizer is not necessarily superior to conventional parameter tuning, and gains in retrieval metrics should not be assumed to produce gains in clinical safety.

A particularly important distinction is between optimizing a system for average performance and ensuring acceptable performance in high-risk cases. In clinical applications, rare but consequential errors may be more important than small improvements in overall accuracy. Consequently, optimization should incorporate safety constraints, evaluation on

challenging cases, and expert assessment of clinically significant errors. Systems should also retain the ability to abstain when reliable evidence cannot be retrieved.

Future investigations should therefore focus on reproducible comparative experiments, clinically meaningful evaluation, and transparent reporting. The most promising approach is likely to combine carefully selected optimization methods with authoritative knowledge sources, evidence verification, and appropriate clinical oversight.

II. CONCLUSION

Metaheuristic optimization presents a promising research direction for improving retrieval accuracy and response reliability in healthcare Retrieval-Augmented Generation systems. Algorithms such as genetic algorithms, particle swarm optimization, ant colony optimization, differential evolution, and simulated annealing can potentially optimize retrieval parameters, document chunking, ranking weights, query reformulation, and evidence-selection strategies. Multi-objective and hybrid optimization approaches may further help balance retrieval relevance, answer grounding, completeness, latency, and computational cost.

However, these potential benefits should not be interpreted as proof that metaheuristic methods consistently outperform conventional optimization techniques in healthcare RAG. Their effectiveness depends on the quality of the knowledge base, the design of the fitness function, the characteristics of the clinical task, and the reliability of the evaluation procedure. Improving document retrieval alone cannot guarantee factual accuracy, appropriate uncertainty handling, or patient safety.

A dependable healthcare RAG system should combine optimized retrieval with authoritative medical evidence, claim-level verification, citation accuracy, temporal validity, privacy safeguards, and expert evaluation. Future research should compare metaheuristic algorithms against strong baseline methods using independent datasets, reproducible experiments, and clinically meaningful safety measures. Ultimately, the goal should not merely be to produce more relevant search results or fluent answers, but to develop transparent, evidence-grounded, computationally efficient, and clinically responsible AI systems that support healthcare professionals without replacing their judgment.

REFERENCES

1. Amugongo, L. M., et al. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*.
2. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
3. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. (2025). *Journal of the American Medical Informatics Association*, 32(4), 605 onward.
4. Kell, G., Roberts, A., Umansky, S., Qian, L., and Marshall, I. J. (2024). Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4), 1009–1024.
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. Rajwar, K., Deep, K., and Das, S. (2023). An exhaustive review of the metaheuristic algorithms for search and optimization: Taxonomy, applications, and open challenges. *Artificial Intelligence Review*, 56, 13187–13257.