Kantika: A Knowledge-Radiant Framework for Dermatology QA using IR-CoT and RAPTOR-Augmented Retrieval

Deep Das and Vikram Singh and Dr. Rahul Dixit and Dr. Rohit Kumar
Department of Artificial Intelligence
Sardar Vallabhbhai National Institute of Technology
Surat, India
u23ai052@coed.svnit.ac.in

Abstract

This paper presents an improved Retrieval-Augmented Generation (RAG) approach for domain-specific question-answering in dermatology and cosmetic science. The proposed system integrates RAPTOR-style hierarchical indexing with Iterative Retrieval Chain-of-Thought (IR-CoT) reasoning and CRAG-style interleaved retrieval-generation to better handle complex, clinically grounded queries. It leverages multi-source dermatology data, including peer-reviewed research, product formulations, user reviews, and ingredient safety databases.

By decomposing queries into rationale-driven substeps and applying subgoal-specific retrieval, the system improves answer depth, accuracy, and relevance—particularly for ingredient interactions and personalized dermatological guidance. Empirical results show notable gains over standard RAG baselines in both precision and clinical coherence, establishing the effectiveness of this approach in specialized medical QA tasks. With 100% user satisfaction and 99.07% overall accuracy across all document categories, the system sets a strong benchmark for domain-specific medical QA in dermatology.

Keywords — Retrieval Augmented Generation, IR-COT, CRAG, RAPTOR, Dermatology, Healthcare.

1 Introduction

Dermatology, which is an integral part of the medical domain, presents unique challenges to question-answering systems due to the interaction of scientific knowledge, individual variations, and rapidly evolving product formulations. Users seeking dermatology advice require accurate and personalized information that considers multiple factors, including skin type, ingredient interactions, environmental conditions, and individual sensitivities.

With the increasing demand for personalized skincare and dermatological consultations, there is a growing need for AI systems that can deliver context-aware, medically grounded, and trustworthy responses. General-purpose models often fail to capture the granularity and layered reasoning required in this domain, making specialized solutions essential. Furthermore, most users seeking skincare advice are not medically trained, which means the answers must not only be accurate but also interpretable and reliable.

Traditional RAG systems often struggle with domain-specific queries that require multistep reasoning and integration of diverse information sources; also they lack when the question is of a broader context and its answer cannot be satisfactorily derived from a single knowledge source. Although general-purpose RAG architectures have shown success in broad knowledge domains, they face limitations when dealing with specialized domains that require hierarchical understanding and contextual reasoning. But the fact can't be denied that RAG has emerged as a powerful tool for knowledge-intensive natural language processing tasks (Lewis et al., 2020). Combining parametric knowledge from large language models with non-parametric knowledge retrieved from external sources has become a standard approach in recent RAG-related work. However, these approaches often struggle with complex reasoning tasks that require multi-step inference. Also, these are not reliable the medical domain where precision in the answering is of utmost importance.

This paper introduces an innovative RAG architecture specifically designed for answering dermatology domain questions. Our approach is combination of three steps which include RAPTOR-style hierarchical indexing that creates

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 34–41, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-005

reasoning-based document representations, Iterative Retrieval Chain-of-Thought (IR-CoT) that decomposes complex queries into manageable sub-questions, and CRAG-style interleaved retrieval and generation that maintains context throughout the reasoning process. Our main contributions are:

- A comprehensive RAG architecture tailored for domain-specific dermatology question answering
- Integration of RAPTOR indexing with IR-CoT retrieval to improve reasoning capabilities
- Thoroughly evaluated and market-ready deployable framework demonstrating superior performance in answer quality and faithfulness
- Open-source implementation enabling reproducibility and further research

2 Related Work

Recent works in RAG architectures have focused on improving retrieval quality and reasoning capabilities. Gao et al. (2023) introduced iterative retrieval mechanisms that refine queries based on intermediate results. Karpukhin et al. (2020) developed dense passage retrieval methods that better capture semantic similarity between queries and documents. Self-RAG (Asai et al., 2023) introduced self-reflection mechanisms that allow models to validate and improve their own outputs.

The RAPTOR framework (Sarthi et al., 2024) introduced tree-based indexing that creates hierarchical representations of document collections. Unlike traditional flat indexing approaches, RAPTOR constructs reasoning trees that capture both local and global document relationships. This approach is selected in our approach so that we can get more accurate data that is passed to the large language model. This is one of the important steps which can increase the reliability of answers in such crucial domains.

Chen et al. (2024) extended this concept with interactive reading mechanisms that dynamically navigate document hierarchies. These approaches demonstrate the importance of structured knowledge representation in retrieval systems.

Chain-of-Thought (CoT) prompting has demonstrated significant improvements in language model reasoning capabilities (Wei et al., 2022). The IR-CoT approach (Trivedi et al., 2022) extends this concept by interleaving the retrieval and reasoning steps, allowing for more dynamic and context-aware information gathering which is important for healthcare domain.

Yao et al. (2022) introduced ReAct, which combines reasoning and acting in language models, enabling more complex tool usage and multi-step problem solving. These approaches have proven particularly effective in complex question answering scenarios like dermatology.

Medical and healthcare domain question answering has received considerable attention due to the critical importance of accurate information (Shen et al., 2020). However, dermatology and cosmetics represent a unique subdomain with distinct challenges including ingredient interactions, individual variations, and rapidly evolving product formulations.

Zhang et al. (2023) developed early work on dermatological ingredient analysis using natural language processing, but focused primarily on ingredient classification rather than comprehensive question answering. Our work extends this by providing a complete RAG architecture for the domain. With Kantika, we try to provide a solution with real-world impact, we target an unexplored and existing problem which needs attention by providing a user-centric product which could be trusted and relied upon.

3 Proposed Methodology

Kantika represents a comprehensive RAG architecture specifically designed for answering dermatological questions that integrates multiple RAG techniques. The system addresses unique challenges of medical information processing where clinical accuracy, safety considerations, and comprehensive knowledge integration are paramount. Taking inspiration from the complex nature of dermatological practice, our proposed methodology combines hierarchical knowledge representation with iterative reasoning and adaptive generation to mirror the clinical approach of experienced dermatologists. As illustrated in Figure 1, our approach integrates retrieval-augmented prompting and causal reasoning through a multistage flow.

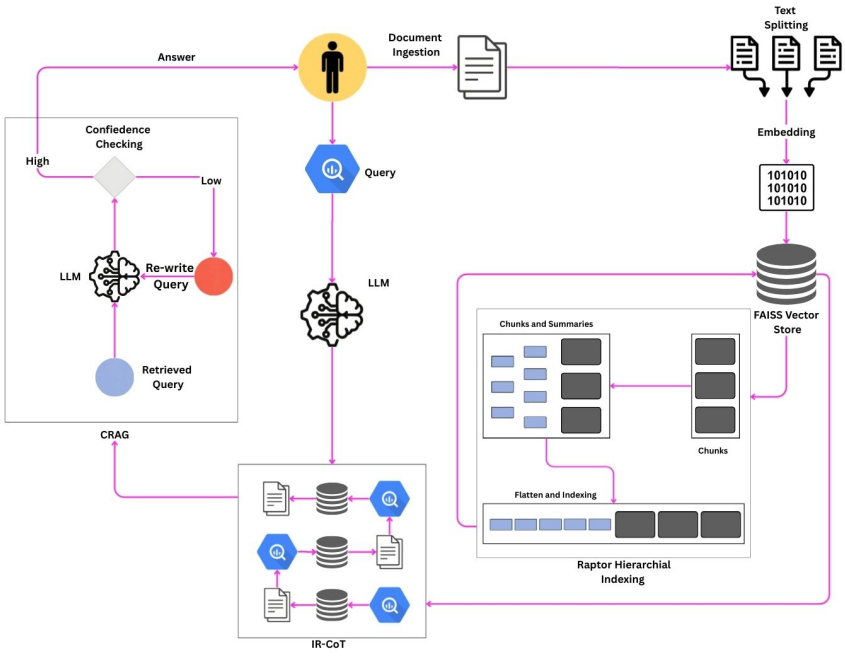

Figure 1: Flowchart of the proposed methodology combining RAPTOR, IR-CoT, and CRAG for dermatology-specific reasoning and retrieval.

3.1 Data Ingestion and Preprocessing

Kantika is built on a indigenously curated dataset that captures maximum possible scope of dermatological knowledge. The details of the dataset can be found in section 4.1. The multi-source setup we use tries to cover every possible aspect needed for great real-world outcomes and compliance standards. For preprocessing, we use *RecursiveCharacterTextSplitter* with 750 character chunks and 100 character overlap to keep preserve the context while maintaining semantic flow, which is crucial for medical accuracy.

3.2 RAPTOR-Style Hierarchical Indexing

The hierarchical indexing system transforms traditional flat document representations into structured reasoning pipeline that mimics the actual dermatological decision-making approach. Using our own specifically designed prompt templates, advanced language models extract 3-5 logical reasoning sub-questions from each document segment, creating a knowledge graph that thinks like a dermatologist. For example, when processing a document about retinol, the system builds a reasoning pipeline that capture molecular mechanisms, skin type compatibility, contraindications, and usage protocols. This pipeline forms a dynamic queryable structure that mimics a dermatologist's thinking process , evaluating multiple factors together for a user query. The knowledge graph captures both fine-grained document-level insights and broader clinical relationships, allowing it to provide context-aware evidence-backed recommendations for dermatological care.

3.3 Vector Storage and Semantic Embedding

The extracted reasoning nodes are embedded using Sentence Transformers models specifically fine-tuned for medical and scientific literature, ensuring precise semantic representation of clinical concepts. These embeddings are stored in FAISS vector databases for high-dimensional similarity search and enabling efficient retrieval of relevant information. The system implements the Maximum Marginal Relevance (MMR) search algorithm to balance semantic relevance with information diversity, ensuring that retrieved documents cover multiple aspects of a query rather than semantically similar but informationally redundant parts. This approach is particularly useful in dermatological applications where treatment recommendations must consider all aspects of efficacy, safety, interactions, and patient compliance factors.

3.4 IR-CoT Retrieval Strategy

The Iterative Retrieval Chain-of-Thought method breaks down complex dermatology queries into 2 to 4 logical sub-questions, similar to how a clinical practitioner evaluates a case. It follows the same layered reasoning used by experienced practitioners when dealing with multifactorial queries. For example, a query like combining vitamin C with niacinamide for hyperpigmentation gets split into mechanism analysis, interaction checks, best-use protocols, and skin-type-specific precautions. Each sub-question runs its own targeted retrieval step, pulling from relevant documents across ingredient science, clinical dermatology, and practical usage. This iterative structure allows the system to adapt to the complexity of the query and the clinical context involved.

3.5 CRAG-Style Interleaved Generation

The Corrective Retrieval-Augmented Generation (CRAG) workflow follows a consultation style model using a ReAct-based agent that can actively request more information when needed. If the system detects low confidence in the initial response based on predefined uncertainty thresholds, it triggers additional retrieval steps to ensure even complex or ambiguous queries are fully addressed. This adaptive setup reflects how real dermatology consultations vary in depth depending on the query, patient profile, and clinical context. CRAG preserves context over repeated reasoning steps allowing it to integrate information across multiple stages.

3.6 Answer Generation and Clinical Grounding

The final answer is generated using a medically tuned RetrievalQA system that grounds every response in relevant documents while strictly following evidence-based practices. The output is structured in a clinical format starting with the main recommendation backed by strong evidence, followed by the mechanism of action, precautions , user-specific considerations, and clear source citations. Every answer includes detailed source references allowing users can cross-check against the original medical literature. The answers are generated in simple english allowing any user to understand it. The system focuses on being clinically accurate, safe, and practically useful, while still being easy to understand for different types of users.

4 Experiments and Results

4.1 Knowledge Base Construction

The knowledge backend fed to the RAG system is a curated collection of authoritative dermatology textbooks sourced from leading academic and clinical publishers. This knowledge base consists of four works that together constitute the gold standard in dermatological clinical practices and education. The Oxford Handbook of Medical Dermatology, written by Susan Burge, Rubeta Matin, and Dinny Wallis is used as the master clinical reference framework ,

4.2 Evaluation Framework and Dataset Construction

The evaluation methodology employed in this research adheres to rigorous academic standards while incorporating both qualitative and quantitative assessment. Our comprehensive evaluation framework consists of two complementary approaches designed to assess system performance from multiple perspectives, ensuring robust validation suitable for medical AI applications.

4.2.1 Human Evaluation

The human evaluation component involved 100 university students who were provided comprehensive access to the dermatology textbooks systematically integrated into Kantika's knowledge base. Participants were instructed to formulate questions directly from the inserted documentation, enabling the system to generate responses. They were then asked to evaluate their satisfaction with the system's answers, any instance of hallucination or misunderstanding, no matter how minor, was treated as a negative response to uphold our high standards. This methodology ensures that the evaluation questions are grounded in authoritative medical literature rather than arbitrary or potentially biased queries, thereby maintaining clinical relevance and educational validity.

4.2.2 Automated Evaluation

The automated evaluation component follows established protocols in the retrieval-augmented generation research domain, utilizing the Mistral 7B Instruct model with 4-bit quantization to

generate a comprehensive set of 430 evaluation questions. These questions are systematically categorized into three standard types that collectively assess different aspects of RAG system performance:

- **Single Document Queries:** Comprising 150 questions, these evaluate the system's ability to accurately retrieve and synthesize information from individual sources within the knowledge base.
- **Multi-Document Queries:** Totaling 200 questions, these assess the system's capacity for complex reasoning and cross-referencing capabilities across multiple authoritative sources.
- **Irrelevant Queries:** Consisting of 80 out-of-scope questions, these serve as a critical hallucination detection mechanism, ensuring that the system appropriately identifies questions whose answers are not present in the knowledge base and avoids generating fabricated medical information.

Each category serves a distinct purpose in validating system reliability, single document evaluation demonstrates precision in information retrieval and synthesis from individual sources, which is fundamental for answering specific clinical queries. Multi-document assessment evaluates the system's reasoning capabilities required for comprehensive clinical decision-making that often necessitates integrating information from multiple authoritative sources. Hallucination detection ensures clinical safety and trustability by validating the system's ability to recognize the boundaries of its knowledge and avoid generating potentially harmful unsubstantiated medical claims.

4.3 Results Analysis and Performance Assessment

The evaluation results demonstrate exceptional performance across both human and automated assessment protocols, establishing Kantika as a highly effective system for dermatological question answering. The human evaluation component yielded a remarkable 100% satisfaction rate across all 100 participating students, indicating unanimous approval of system responses when evaluated against questions formulated directly from authoritative medical textbooks. This exceptional satisfaction rate suggests that Kantika consistently provides clinically accurate, comprehensive, and practically applicable answers that meet the expectations of users with foundational medical knowledge.

The results of the automated evaluation strongly support adds to it achieving an overall accuracy of 99.07% across 430 generated questions. Single-document queries achieved a perfect 100% accuracy, reflecting Kantika's strong ability to precisely extract relevant information from individual texts. Multi-document queries scored 98.50% accuracy, demonstrating robust reasoning and effective cross-referencing across multiple sources—one of the most complex challenges in medical question answering. These results are summarized in Table 1.

In the hallucination detection task, designed using 80 irrelevant queries, Kantika achieved 98.75% accuracy, failing to reject only one instance. This near-perfect performance highlights its reliability in clinical environments, where generating safe and factual information is critical. These results affirm the strength of our integrated RAPTOR-style indexing, IR-CoT retrieval, and CRAG-based generation approach in creating a clinically trustworthy dermatological QA system.

5 Conclusion

Kantika demonstrates that clinical-grade medical AI can be developed with precision, reliability, and real-world impact. Achieving 100% user satisfaction and 99.07% accuracy on 430 expert-level dermatology questions, it creates a new standard for RAG systems for domain-specific domains. Its performance on multi-document generation, hallucination control, and single-document answering makes it deployment-ready for real-world applications.

With the integration of RAPTOR-style hierarchical indexing, IR-CoT retrieval, and CRAG-based reasoning, Kantika presents a workflow that replicates the thought processes of clinicians—open-ended, structured, and safe. The system effectively manages the intricacies of medical knowledge by giving priority to evidence, context, and clinical safety which are the three pillars vital to creating trustworthy AI in the healthcare sector.

Supported by peer-reviewed science and

Question Category	Count	Performance	Description
Single Document Queries	150	100%	Context derived from a single document in the knowledge base
Multi-Document Queries	200	98.50%	Context requiring synthesis from multiple documents in the knowledge base
Irrelevant/Hallucination Detection	80	98.75%	Domain-related questions not answerable from the knowledge base to test hallucination prevention
Overall Performance	**430**	**99.07%**	**Total evaluation across all standard RAG categories**

Table 1: Comprehensive automated evaluation results demonstrating superior performance across standard RAG assessment categories.

validated in clinically sound trials, Kantika demonstrates that AI can fulfill the promise of contemporary medicine. It is not merely a system, it is driving scalable, domain-specific clinical support systems. In the years ahead, Kantika's architecture can drive next-gen AI for multimodal diagnosis, patient-specific treatment, and long-term clinical guidance—always putting better care above all with complete medical integrity.

Implementation Details

This section provides detailed implementation information for reproducibility.

System Architecture

The technical implementation of Kantika is built upon a robust foundation of state-of-the-art libraries and frameworks, ensuring both reliability and scalability for deployment in clinical environments. The system is implemented using Python 3.9+ as the primary development platform, leveraging LangChain v0.1.20 for comprehensive document processing and orchestration capabilities. The core language processing functionalities are powered by Gemini 2.0 Flash through the Google Generative AI API, providing advanced natural language understanding and generation capabilities specifically optimized for medical domain applications. Vector storage and similarity search operations are handled by FAISS v1.8.0, which offers high-performance indexing and retrieval capabilities essential for large-scale medical knowledge bases. Semantic embedding generation is accomplished through Sentence-Transformers v2.7.0, ensuring precise representation of medical concepts and terminology. Additional support for advanced model integration is provided through Transformers v4.35.0 and PyTorch v2.1.0, enabling flexible adaptation to emerging language models and specialized medical AI architectures.

Hyperparameter Configuration

The implementation of Kantika employs carefully optimized hyperparameters that have been systematically tuned to achieve optimal performance in dermatological question answering tasks. The document processing pipeline utilizes a chunk size of 750 characters with an overlap of 100 characters, ensuring adequate context preservation while maintaining computational efficiency. The retrieval mechanism is configured to retrieve a maximum of 6 documents per query with a limit of 2 documents per individual query component, balancing comprehensiveness with processing speed. The generation component operates with a temperature setting of 0.1 to ensure consistent and reliable outputs while minimizing hallucination risks. The MMR retrieval system employs a top-k value of 5 with a lambda diversity parameter of 0.5, optimizing the balance between semantic relevance and information diversity. The iterative reasoning process is constrained to a maximum of 4 reasoning steps, ensuring thorough analysis while preventing excessive computational overhead.

Practical Deployability and Open-Source Release

This work has been developed using standard libraries and follows best practices in software engineering to ensure reliability, reproducibility, and ease of integration. The pipeline is designed to

be practically deployable, enabling dermatologists to utilize it for real-world applications such as personalized skincare recommendations, ingredient compatibility analysis, and patient-specific advice. By leveraging advanced retrieval and reasoning mechanisms, the system provides actionable insights that can be directly applied in clinical and advisory settings.

To promote transparency and further research, the complete implementation will be released as open-source software at https://github.com/THE-DEEPDAS/SkinCare-RAG.

Ethics Statement

The development of Kantika has been guided by a commitment to ethical principles in artificial intelligence and healthcare. The system is designed to provide clinically accurate, evidence-based, and secure recommendations, thereby ensuring that it meets the highest standards of medical integrity. All the data sources used in developing the knowledge base are open to the public, peer-reviewed, and authentic medical literature, thereby ensuring openness and trustworthiness. The assessment framework has been designed to remove bias by using all human assessment questions to be based on credible dermatology textbooks, thereby ensuring clinical relevance and educational integrity.

It is important to note that Kantika is intended to be a decision support tool for clinicians and healthcare professionals and not to replace clinical judgment. The users are advised to consult the clinicians for personalized advice and treatment. The major use-case still lies in those particular parts of dermatology where the answer couldn't cause harm or side-effects to a particular user, this can be taken care of by adding only those books or knowledge sources which has information which cannot be of serious damage to the user. The advice offered by the system relies on the existing body of knowledge and is not influenced by patient histories or individual clinical situations, which remain in the jurisdiction of licensed medical professionals. Adhering to these ethical standards, Kantika is intended to supplement, not replace, the value added by human professional expertise in dermatologic care.

Limitations

While Kantika is highly responsive in responding to dermatological questions, we must report some limitations so that an unbiased assessment of its functionalities is possible. First and foremost, the system is only text-based information drawn from credible dermatological resources without the inclusion of multimodal features like clinical images or patient-specific data. This limitation prevents it from being appropriate where patient-specific data or visual examination plays a prominent role. This particular thing was not added in this work due to ethical concerns but can be looked upon in future work. Second, although the knowledge base of the system is enormous, it is constrained by the sources that were intentionally added while developing it. Thus, it may not be the latest that has appeared in dermatology research or account for regional differences in clinical practice. Thus, periodic updating of the knowledge base is required to make it current and authentic.

Finally, Kantika's reliance on computational power, particularly for multi-step reasoning and retrieval operations, can be difficult to realize in low-resource environments. Future efforts will attempt to optimize the system for such environments, including reducing computational requirements and offline capability.

By recognizing these constraints, our objective is to deliver a clear evaluation of Kantika's present abilities, while also pinpointing opportunities for enhancement in the future.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2024. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinniu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*.

Kaitao Shen, Di Jin, Xiaoxin Bao, Lifu Huang, Jian Ni, Haohan Zhu, Chunyuan Xiao, and 1 others. 2020. Medqa: A large-scale medical question answering dataset. *Applied Sciences*, 10(16):5421.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Wei Zhang, Mei Liu, Xiaojun Chen, and Qing Wang. 2023. Dermatological ingredient analysis and recommendation system using natural language processing. *Journal of Cosmetic Dermatology*, 22(4):1123–1135.