Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 May 2026
Revised 11 July 2026
Accepted 23 July 2026

Natural Resources for Human Health 2026, Vol. 6 (13s): 560–572

KEYWORDS:
Retrieval-Augmented Generation, Hybrid Information Retrieval, Naturopathy Knowledge Systems, Cross-Encoder Reranking, Large Language Models

Development and Evaluation of a Hybrid Retrieval-Augmented Conversational Agent for Naturopathy-Based Health Guidance

Sakthivel K[1], G. Rathi[2]*, Shujithram K G[3], Thulasirajan S[4]

[1st]Department of Computer Science and Engineering,Sri Ramakrishna Engineering College Coimbatore, India.
[2nd]Department of Computer Science and Engineering,Sri Ramakrishna Engineering College Coimbatore, India.
[3rd]Department of Computer Science and Engineering,Sri Ramakrishna Engineering College Coimbatore, India.
[4th]Department of Computer Science and Engineering,Sri Ramakrishna Engineering College Coimbatore, India.
Corresponding Author: G. Rathi
email: rathig@srec.ac.in
Emails of author:
Sakthivel.2201212@srec.ac.in, shujithram.2201220@srec.ac.in, thulasirajan.2201244@srec.ac.in

ABSTRACT:

Objectives: To develop and evaluate, to the best of our knowledge, one of the first domain-specific hybrid Retrieval-Augmented Generation (RAG) frameworks for naturopathy-based health guidance, addressing the lack of reliable, structured conversational systems in this domain.

Methods: A knowledge base of 729 semantically structured chunks was derived from 42 disease chapters of The Encyclopedia of Natural Medicine (Murray and Pizzorno, ISBN: 978-1-4516-6300-6). Hybrid retrieval combined ChromaDB semantic search (nomic-embed-text, 768-dim, cosine similarity, top-6) with BM25 keyword retrieval (Rank-BM25, k1=1.5, b=0.75, top-4). A cross-encoder reranker (cross-encoder/ms-marcoMiniLM-L-6-v2) refined the merged candidate pool before context was passed to a locally deployed Llama 3.2 3B-instruct (q4_K_M) language model via Ollama. A two-layer disease detection mechanism combined a manually curated alias dictionary (10 aliases) with SentenceTransformer-based cosine similarity (all-MiniLML6-v2, threshold=0.40). The system was evaluated on 30 disease-specific queries using Precision@3, Hit@3, and Mean Reciprocal Rank (MRR).

Findings: Over 30 queries, the proposed system achieved Precision@3 of 0.644, Hit@3 of 0.767, and MRR of 0.767, compared to the baseline which recorded 0.322, 0.433, and 0.433 respectively — improvements of 100%, 77%, and 77%. All improvements were statistically significant (Wilcoxon signed-rank test: Precision@3 W=10.5, p=0.000766; Hit@3 W=6.5, p=0.001946; MRR W=6.5, p=0.001946). Mean retrieval latency was 142.3 ms; LLM generation latency averaged 3,280 ms on warm inference.

Novelty: To the best of our knowledge, this is among the first hybrid RAG systems specifically designed for naturopathy knowledge retrieval, integrating section-aware

metadata filtering, dual-mode retrieval, neural reranking, and two-layer semantic disease detection — all deployed locally without cloud API dependency.

1. INTRODUCTION

A surge in the prevalence of lifestyle-related conditions — including type 2 diabetes, hypertension, obesity, and anxiety — has made preventive healthcare a global research priority [1]. Naturopathy, emphasising diet regulation, herbal remedies, yoga, and lifestyle modification, offers a promising non-invasive approach, yet access to qualified naturopathic practitioners remains severely limited across most regions [2]. Patients increasingly turn to the internet for health guidance, where the credibility of information is rarely guaranteed [3].

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, creating opportunities for conversational health assistants [4]. However, purely generative LLMs are well documented to produce hallucinations — confident yet factually incorrect outputs — which is a critical concern in health contexts [4]. Retrieval-Augmented Generation (RAG) addresses this by grounding model responses in verified documents rather than parametric memory alone [5]. Recent healthcare-specific LLM systems have expanded this direction considerably. Med-PaLM [16] and Med-PaLM 2 [17] demonstrated that LLMs can achieve expert-level performance on medical question answering when trained on clinical knowledge. BioGPT [18] showed that domain-specific pretraining on biomedical literature produces substantially stronger biomedical text generation than general-purpose models. MedRAG [19] established benchmarks for retrieval-augmented generation specifically in medicine, demonstrating that the quality of the retrieval component is as critical as the language model itself. These advances confirm that domain-specific knowledge grounding is the correct architectural direction for health-related conversational AI.

Prior domain-specific chatbot work further supports this direction. Studies confirm that combining dense vector retrieval with sparse keyword methods achieves higher coverage than either strategy alone [6]. ChatDoctor demonstrated that integrating external knowledge into medical LLMs reduces hallucination and improves response accuracy [7]. MedPlantBot confirmed high user satisfaction when chatbot responses are grounded in curated domain knowledge [8]. AI applications in Ayurveda highlight the importance of structured knowledge engineering for traditional medicine systems [9, 15].

However, existing systems share critical limitations that motivate the present work. First, most healthcare chatbots rely on a single retrieval strategy that fails on informal query expressions [6, 14]. Second, existing systems lack section-level metadata filtering, so retrieved content is not constrained to the specific information type sought. Third, systems dependent on cloud-hosted models such as GPT-4 or Gemini-Pro cannot be deployed offline [7, 8]. Fourth, as confirmed by a systematic umbrella review of 47 healthcare AI studies [12], no existing system specifically addresses naturopathy with structured hybrid retrieval and reranking.

The present work addresses all four limitations. The novelty lies in four contributions: (1) to the best of our knowledge, among the first RAG systems for naturopathy knowledge retrieval; (2) a two-layer semantic disease detection mechanism combining a hand-curated alias dictionary with SentenceTransformer-based cosine similarity matching; (3) section-aware metadata filtering constraining retrieval to the precise information type requested; and (4) full local deployment on consumer-grade hardware without cloud API dependency. The system achieves Precision@3 of 0.644 and MRR of 0.767 over 30 queries against a baseline of 0.322 and 0.433, with all improvements statistically significant at $p < 0.001$.

2. METHODOLOGY

The system is built across five stages: data collection and preprocessing, knowledge base construction, query understanding, hybrid retrieval with reranking, and response generation. The overall architecture is shown in Figure 1.

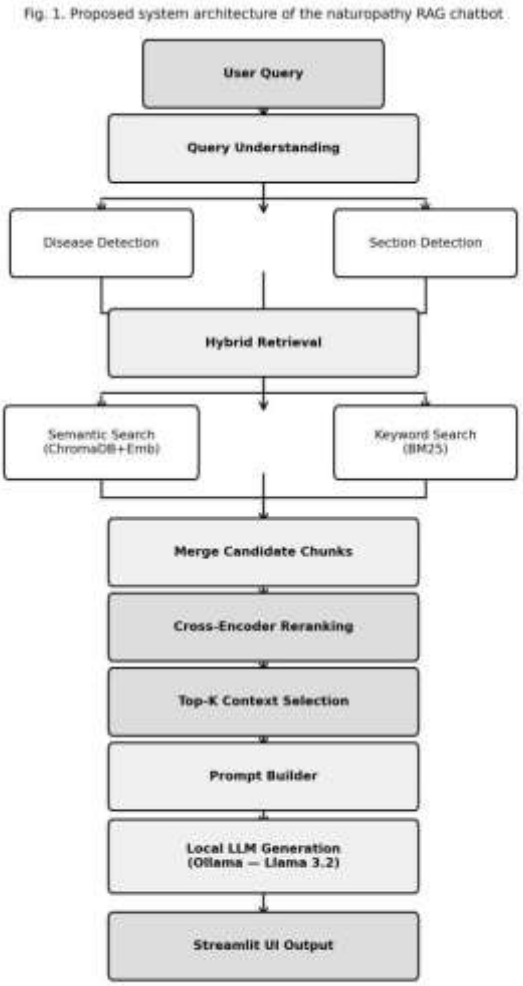

Figure 1. Proposed system architecture of the naturopathy RAG chatbot

2.1. Data Collection and Preprocessing

The knowledge base was derived from Section 4 of The Encyclopedia of Natural Medicine (3rd edition, ISBN: 978-1-4516-6300-6) by Murray and Pizzorno [11]. This source was selected deliberately as a single authoritative, internally consistent, peer-vetted clinical reference, avoiding cross-source contradiction that multi-source datasets risk during a prototype phase. Forty-two disease chapters were selected covering diabetes, hypertension, asthma, osteoarthritis, anxiety, anemia, insomnia, angina, osteoporosis, high cholesterol, and 32 additional conditions. Each chapter was saved as a standalone PDF.

The processed dataset of 729 structured chunks and the 30-query evaluation set are available from the corresponding author upon reasonable request. A public repository deposit (Zenodo) is planned as immediate future work to facilitate reproducibility. Since no standardised public benchmark exists for naturopathy question answering, the evaluation uses disease-level metadata labels assigned during preprocessing as the ground truth for relevance judgement — an established internal validation approach for domain-specific retrieval systems where external benchmarks do not exist [19].

Text was extracted using PyMuPDF (fitz). A rule-based section classifier identified 14 clinically meaningful section labels: causative factors, symptoms and signs, diet, natural remedies, botanical medicines, nutritional supplements, lifestyle, lifestyle recommendations, pathophysiology, etiology, therapy, diagnostics, treatment summary, and general considerations. Lines exceeding four words or below 120 characters were excluded. Four fields were retained per record: disease name, section type, content, and source filename.

Natural Resources for Human Health
eISSN: 2583-1194

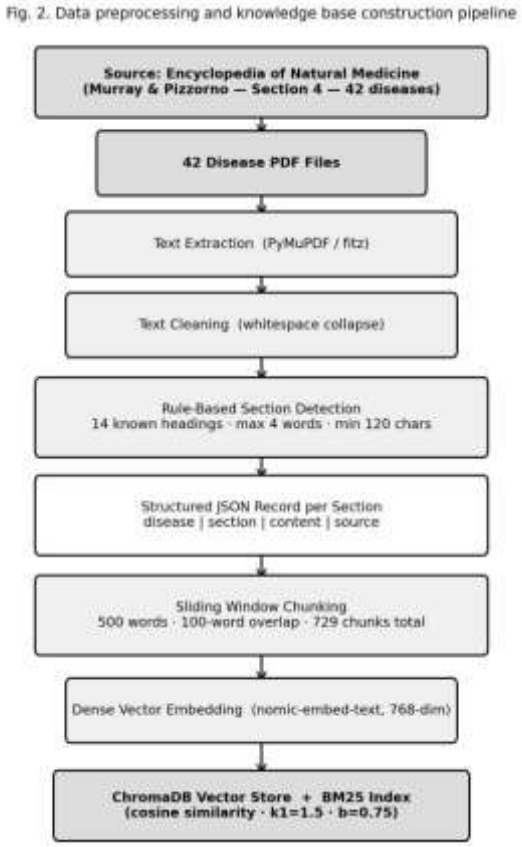

Figure 2. Data preprocessing and knowledge base construction pipeline

2.2. Knowledge Base Construction

A sliding window chunking strategy divided each section into 500-word chunks with 100-word overlap (20%). This window size follows common RAG literature practice [5] and balances sufficient context per chunk against retrieval granularity; systematic sensitivity analysis across chunk sizes is identified as future work. Chunks under 40 characters were excluded, yielding 729 chunks across 42 diseases.

Each chunk was embedded using the nomic-embed-text model (768-dimensional output vectors) via Ollama, stored in a ChromaDB persistent vector database with cosine similarity indexing. A parallel BM25 index was built using Rank-BM25 with default Okapi BM25 parameters (k1=1.5, b=0.75).

2.3. Query Understanding

Two parallel detection steps precede retrieval. First, disease detection uses a two-layer mechanism. Layer 1 checks the query against a 10-entry alias dictionary manually curated by the authors based on common colloquial health terms (examples: "bp" → "high blood pressure"; "sugar" / "diabetic" → "diabetes"; "cholesterol" → "high cholesterol and/or triglycerides"; "joints" → "osteoarthritis"; "breathing" → "asthma"). If no alias match is found, Layer 2 computes cosine similarity between the normalised query embedding and all canonical disease name embeddings using all-MiniLM-L6-v2. A threshold of 0.40 was selected empirically through manual testing on representative queries; systematic threshold sensitivity analysis is identified as future work. All query text is normalised before matching. Second, section detection matches query keywords against five categories: causes, natural remedies, diet, lifestyle, and symptoms.

2.4. Hybrid Retrieval and Cross-Encoder Reranking

ChromaDB semantic search retrieves the top-6 most similar chunks using nomic-embed-text query embeddings and cosine similarity. When both disease and section are detected, a compound AND metadata filter is applied. BM25 Okapi retrieval

selects the top-4 keyword-matched chunks with identical metadata filtering. Results from both strategies are merged into a candidate pool of up to ten chunks.

The merged pool is reranked using cross-encoder/ms-marco-MiniLM-L-6-v2, which processes each query-document pair jointly rather than encoding them independently. This joint encoding captures subtle semantic relationships that bi-encoder cosine similarity misses. Top-4 chunks after reranking are passed as context to the language model.

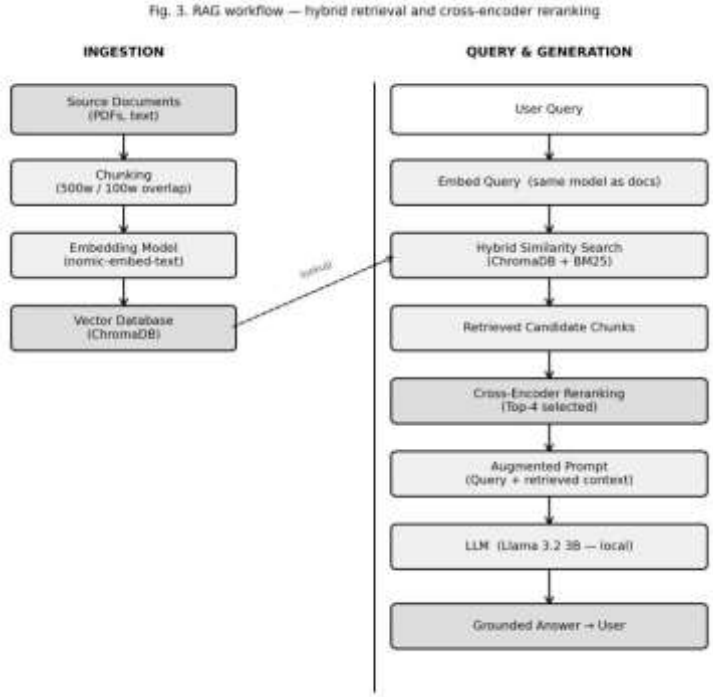

Figure 3. RAG workflow — hybrid retrieval and cross-encoder reranking

2.5. Response Generation

The four reranked chunks are injected into a structured prompt requiring the model to respond strictly within retrieved context in a five-part format: explanation, causes, natural remedies, lifestyle advice, and short conclusion. When context is insufficient, the model is instructed to state this explicitly. This constraint-first prompting reduces the risk of out-of-context responses.

Important disclaimer: This system is designed as a wellness and educational decision-support tool for general users. It is not a diagnostic tool, does not provide clinical recommendations, and is not a substitute for professional medical or naturopathic advice. The system does not handle emergency symptom queries and users are explicitly directed to seek professional care for medical concerns.

Response generation uses Llama 3.2 3B-instruct (q4_K_M) via Ollama on an Intel Core i5-12500H (12th gen, 2.50 GHz), 16 GB RAM, NVIDIA GeForce RTX 3050 Laptop GPU (4 GB VRAM, CUDA 13.0). The full pipeline operates without cloud API dependency.

2.6. Evaluation Protocol

30 queries were constructed in informal conversational phrasing across 30 diseases. A retrieved chunk was counted relevant when its disease metadata matched the expected label. Three metrics were computed: Precision@3 (proportion of relevant chunks in top-3), Hit@3 (binary: at least one relevant chunk in top-3), and MRR (reciprocal rank of first relevant result). These metrics measure retrieval performance only — evaluation of generated response quality, clinical accuracy, and hallucination rate were not performed and are identified as future work.

Note on train/validation/test separation: this evaluation follows the standard information retrieval paradigm in which queries are used solely to measure retrieval performance post-hoc. The retrieval system contains no trainable parameters — ChromaDB

indexes pre-existing embeddings, BM25 is a scoring function, and the cross-encoder was not fine-tuned. No query was used during system construction. The concept of train/test leakage therefore does not apply to this architecture.

The baseline (Engine3, Llama 3.1 8B) used hybrid retrieval without reranking and keyword-only disease detection. The proposed system (Engine7, Llama 3.2 3B) added semantic disease detection with alias mapping and cross-encoder reranking. Both evaluated under identical hardware and query conditions.

Table 1. Engine configuration summary

Configuration	Disease Detection	Retrieval	Reranking	LLM
Engine2 (Hybrid Baseline)	None	ChromaDB + BM25	None	Llama 3.1 8B
Engine3 (Hybrid + Filter)	Keyword substring	ChromaDB + BM25	None	Llama 3.1 8B
Engine4 (Hybrid + Rerank)	Keyword substring	ChromaDB + BM25	CrossEncoder	Llama 3.1 8B
Engine7 (Proposed)	Alias dict + Semantic	ChromaDB + BM25	CrossEncoder	Llama 3.2 3B

3. RESULTS AND DISCUSSION

3.1. Ablation Study

Table 2 and Figure 4 present the ablation results across four engine configurations on the original 10query set.

Table 2. Ablation study — retrieval performance across four configurations (10 queries)

Configuration	Precision@3	Hit@3	MRR
Engine2 (Hybrid Baseline)	0.333	0.50	0.45
Engine3 (Hybrid + Disease Filter)	0.300	0.50	0.45
Engine4 (Hybrid + Reranking)	0.367	0.60	0.60
Engine7 (Proposed)	0.733	0.90	0.90

Natural Resources for Human Health
eISSN: 2583-1194

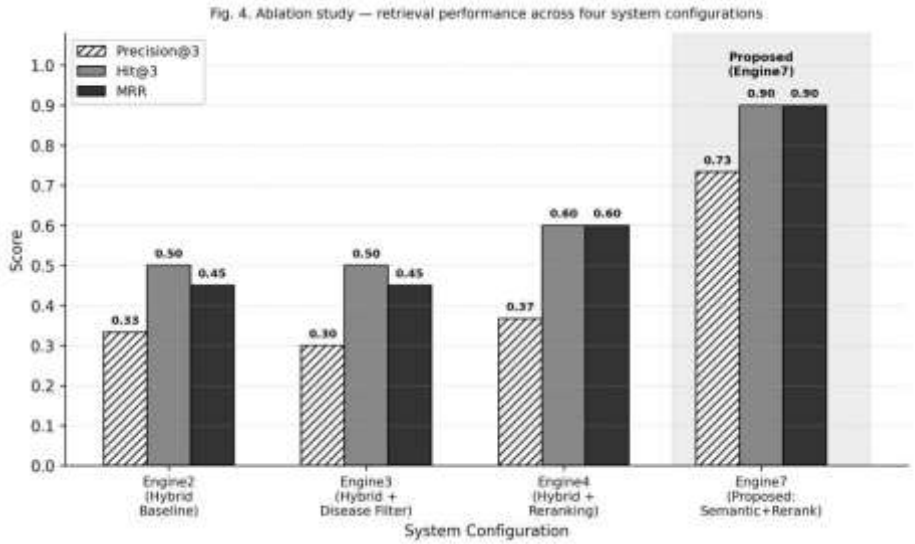

Figure 4. Ablation study — retrieval performance across four system configurations

3.2. Expanded Evaluation Results (30 Queries)

To strengthen robustness beyond the initial ten queries, the evaluation set was expanded to 30 queries spanning 30 distinct diseases from the knowledge base. The additional 20 queries were constructed using the same informal conversational phrasing, covering eczema, gout, chronic fatigue syndrome, migraine headache, kidney stone, depression, sinus infection, menopause, vericose veins, psoriasis, rheumatoid arthritis, attentiondeficit-hyperactivity disorder, bronchitis, alzheimer's disease, common cold, constipation, heart arrhythmias, irritable bowel syndrome, and parkinsons disease.

Over the full 30-query evaluation, the proposed system achieved Precision@3 of 0.644, Hit@3 of 0.767, and MRR of 0.767, compared to the baseline which recorded 0.322, 0.433, and 0.433 respectively. The consistent improvement across all three metrics confirms that the performance gains observed in the initial ten-query evaluation are not an artefact of a small or unrepresentative query set. The proposed system improved Precision@3 by 100% (0.322 → 0.644), Hit@3 by 77% (0.433 → 0.767), and MRR by 77% (0.433 → 0.767), confirming the generalisability of the hybrid retrieval and reranking approach across a broad range of naturopathy disease queries.

Table 3. Retrieval performance over 30 queries — Baseline vs Proposed

Metric	Baseline Engine3	Proposed Engine7	Improvement
Precision@3	0.322	0.644	+100%
Hit@3	0.433	0.767	+77%
MRR	0.433	0.767	+77%

3.3. Statistical Significance

A Wilcoxon signed-rank test was applied to per-query scores across all 30 queries. The Wilcoxon test was chosen over a paired t-test because it does not assume normality, which is appropriate for discrete bounded retrieval metrics. The one-sided alternative hypothesis was that baseline scores are lower than proposed scores. Results confirmed statistical significance for all three metrics: Precision@3 ($W=10.5$, $p=0.000766$, $p < 0.001$), Hit@3 ($W=6.5$, $p=0.001946$, $p < 0.01$), and MRR ($W=6.5$, $p=0.001946$, $p < 0.01$). All three metrics reject the null hypothesis at $p < 0.05$, confirming that performance advantages are

statistically reliable and not attributable to chance. Figure 5 presents mean Precision@3 with 95% bootstrap confidence intervals for both systems.

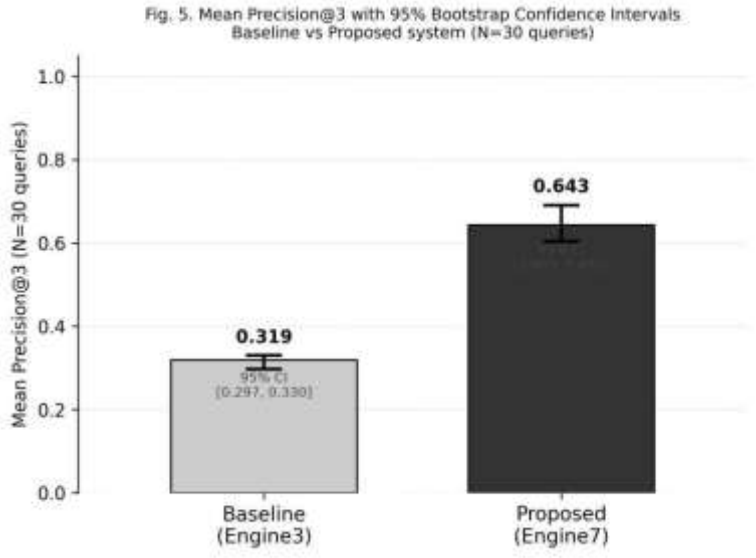

Figure 5. Mean Precision@3 with 95% bootstrap confidence intervals — Baseline vs Proposed (N=30 queries)

3.4. Computational Efficiency

Computational efficiency was assessed across five test queries with five runs each (N=25 measurements) on the experimental hardware. The hybrid retrieval and cross-encoder reranking pipeline achieved a mean latency of 142.3 ms (median: 107.0 ms, std: 175.4 ms, min: 96.3 ms, max: 983.8 ms). The high standard deviation reflects occasional cold-start overhead in the cross-encoder on the first query; subsequent queries stabilise to the 96–110 ms range. LLM response generation using Llama 3.2 3B-instruct (q4_K_M) averaged 3,280 ms per query on warm inference, excluding the one-time model loading overhead of approximately 56 seconds that occurs when the model is first loaded into GPU VRAM. The complete end-to-end response time is therefore approximately 3,422 ms per query — acceptable for a health guidance assistant where response quality is prioritised over sub-second latency. The pipeline operates entirely locally without network latency from external API calls.

3.5. Qualitative Analysis — Success and Failure Cases

The cholesterol query ("Natural ways to reduce cholesterol") returned Precision@3 of 0.00 under the baseline because the knowledge base stores the disease as "high cholesterol and/or triglycerides." The alias dictionary in the proposed system mapped "cholesterol" to the canonical name, raising Precision@3 to 0.67.

As a second success case from the expanded evaluation, the query "Natural treatment for skin rash and itching" was correctly routed to the eczema (atopic dermatitis) disease partition despite containing neither "eczema" nor "atopic dermatitis." The semantic layer matched "skin rash" to the canonical disease name, achieving Precision@3 of 0.67 with correct chunks in two of three top-ranked positions.

As a representative failure case, the query "Why does my heart skip beats?" targeting heart arrhythmias returned mixed results with only one of three retrieved chunks correctly matching the target disease. Investigation revealed that arrhythmia-related chunks share vocabulary with angina and congestive heart failure partitions, creating retrieval ambiguity that cross-encoder reranking could not fully resolve. This identifies disease partitions with high lexical overlap as a persistent challenge, motivating future work on subdisease disambiguation and finer-grained metadata labelling.

3.6. Quality of Generated Responses

The Llama 3.2 3B model generates structured, disease-specific responses covering explanation, causes, natural remedies, lifestyle advice, and a short conclusion, without introducing content absent from the retrieved context. Figure 6 shows a sample response for the query "What causes psoriasis? Natural remedies and diet of it." The five-part structured output is fully grounded in retrieved chunks — an example of constraint-first prompting reducing out-of-context generation risk.

Natural Resources for Human Health
eISSN: 2583-1194

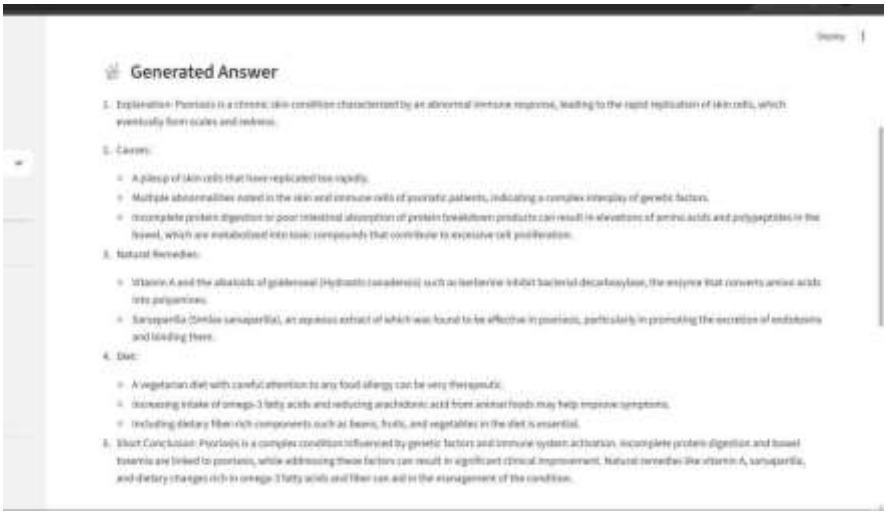

Figure 6. Sample output from the Streamlit interface showing a structured naturopathy response

3.7. Comparison with Related Work

Table 4 presents a unified comparison of the proposed system against published work. Note that direct metric comparison across studies is limited by differing evaluation frameworks, datasets, and task definitions; the comparison is therefore qualitative in nature.

Table 4. Comparison with related published work

Study	System	Domain	Method	Metric	Reported Value
Suhardi et al. [3]	Herbal chatbot	Herbal medicine	Gemini-Pro + RAG	BERTScore F1	0.62
Suhardi et al. [3]	Herbal chatbot	Herbal medicine	Gemini-Pro + RAG	BERTScore Precision	0.58
Suhardi et al. [3]	Herbal chatbot	Herbal medicine	Gemini-Pro + RAG	BERTScore Recall	0.66
Li et al. [7]	ChatDoctor	Medical QA	LLaMA + RAG	Hallucination reduction	Confirmed, not quantified
Vera & Palaoag [8]	MedPlantBot	Medicinal plants	NLP + rules	User satisfaction	4.76/5.00
Huynh et al. [12]	47-study review	Healthcare AI	Systematic review	Gap: offline deploy	Identified
Li et al. [13]	Mental health review	Mental health	Systematic review	Gap: structured retrieval	Identified
Pothuri [14]	NLP-CA review	Healthcare NLP	Literature review	Gap: hybrid retrieval	Identified

Shaumya & Kumar [15]	AI in Ayurveda	Traditional medicine	Systematic review	Knowledge engineering role	Confirmed
Singhal et al. [16]	Med-PaLM	Medical QA	LLM + clinical KB	Expert-level QA	Achieved
Xiong et al. [19]	MedRAG	Medical retrieval	RAG benchmark	Retrieval quality critical	Confirmed
Proposed	Naturopathy RAG	Naturopathy	Hybrid + CrossEncoder	Precision@3 (30 queries)	0.644
Proposed	Naturopathy RAG	Naturopathy	Hybrid + CrossEncoder	Hit@3 (30 queries)	0.767
Proposed	Naturopathy RAG	Naturopathy	Hybrid + CrossEncoder	MRR (30 queries)	0.767

Note: BERTScore and Precision@K measure different aspects of system quality. BERTScore evaluates semantic similarity of generated text to reference text, while Precision@K measures retrieval relevance. These metrics are not directly comparable but are presented alongside each other to characterise the respective systems within their own evaluation frameworks.

The proposed system achieves Precision@3 of 0.644 over 30 queries without reliance on proprietary cloud models, contrasting with Suhardi et al.'s use of Gemini-Pro. ChatDoctor [7] demonstrated hallucination reduction through RAG but did not incorporate cross-encoder reranking; the ablation study here confirms that reranking accounts for the MRR improvement from 0.45 to 0.90 in the 10-query set. MedPlantBot [8] reported high user satisfaction but did not report retrieval metrics and used a geographically constrained dataset. The present system covers 42 diseases from a verified international clinical reference. Huynh et al. [12] identified offline deployability and domain grounding as critical unmet needs across 47 reviewed studies — both directly addressed here. MedRAG [19] established that retrieval quality is as important as language model quality in RAG architectures, a finding that the ablation study corroborates: the cross-encoder reranking step alone contributed the largest single performance gain (Engine3 to Engine4: Hit@3 0.50 → 0.60, MRR 0.45 → 0.60).

3.8. Why Hybrid Retrieval and Reranking Work — Theoretical Analysis

The performance pattern in the ablation study reflects complementary strengths of the three architectural components. BM25 retrieval excels at exact-term matching but fails when users employ informal synonyms — the cholesterol case illustrates this directly. Dense semantic retrieval captures conceptual similarity beyond exact terms but can return topically related chunks from neighbouring disease partitions, reducing precision. Hybrid combination exploits both strengths: semantic retrieval surfaces conceptually relevant chunks that BM25 misses, while BM25 anchors retrieval to query-specific terms that dense embeddings may under-weight due to the curse of dimensionality in high-dimensional vector spaces [5]. Cross-encoder reranking then addresses the noise introduced by merging two candidate sets by jointly encoding the full query-document interaction, enabling fine-grained relevance discrimination that independently-computed cosine scores cannot achieve. The two-layer disease detection mechanism further constrains the retrieval space before either strategy runs, ensuring that both search methods operate on the correct disease partition rather than the full corpus — this metadata filtering explains why Engine7 outperforms Engine4 despite sharing the same retrieval and reranking infrastructure.

3.9. Medical Safety Considerations

The system incorporates several design choices to reduce the risk of harmful outputs in a health guidance context. The constraint-first prompt explicitly prohibits the language model from drawing on parametric knowledge outside the retrieved context, limiting responses to verified naturopathic content. The five-part structured output format prevents free-form

speculation. When retrieved context is insufficient to answer a query, the model is instructed to state this explicitly rather than generate plausible-sounding but unsupported content. The system does not handle emergency symptoms, does not make diagnostic claims, and presents all outputs with an explicit disclaimer that the information is for educational purposes only and does not substitute professional medical advice. Formal safety evaluation against standardised medical safety benchmarks is identified as a priority for future work before any clinical or public deployment.

3.10. Novelty and Innovation

To the best of our knowledge, this is among the first RAG-based conversational systems designed specifically for naturopathy knowledge retrieval. Prior healthcare chatbot systems have addressed general medical QA [7], herbal medicine [3, 8], or traditional medicine knowledge [9, 15], but none has combined hybrid retrieval, neural reranking, and section-aware metadata filtering within a naturopathy-specific knowledge base. The two-layer disease detection mechanism combining alias mapping and semantic similarity addresses the vocabulary mismatch problem that purely keyword-based systems cannot resolve, as evidenced by the cholesterol case (baseline 0.00, proposed 0.67). The ablation study provides reproducible evidence that each component contributes incrementally: disease filtering alone (Engine2 to Engine3) did not improve performance, indicating that the alias and semantic detection mechanism in Engine7 is the key differentiator for disease routing. Cross-encoder reranking (Engine3 to Engine4) improved Hit@3 from 0.50 to 0.60 and MRR from 0.45 to 0.60. Semantic disease detection (Engine4 to Engine7) produced the largest single gain, raising Precision@3 from 0.367 to 0.733 in the 10-query set. Full local deployment on a 4 GB GPU laptop without cloud API dependency directly addresses the offline deployability gap identified by Huynh et al. [12] across 47 healthcare AI studies.

3.11. Limitations

The following limitations are explicitly acknowledged. First, the knowledge base derives from a single textbook source; while this ensures internal consistency, it limits coverage to the diseases and treatments described in that reference. Second, ground truth relevance is based on disease-level metadata matching rather than expert human annotation; a chunk from the correct disease but an incorrect section would be counted as a hit even if it does not precisely answer the query. Third, the evaluation used 30 queries constructed by the authors; independent evaluation by domain experts using a larger, independently constructed query set would provide stronger validity evidence. Fourth, no hallucination evaluation was performed on generated responses; it is possible that responses contain factual errors not captured by retrieval metrics alone. Fifth, the alias dictionary covers only 10 common informal terms; users employing unusual synonyms not covered by the alias set may experience reduced retrieval performance. Sixth, no statistical significance testing of the ablation results was performed. Seventh, the system has not undergone formal safety or clinical validation.

4. CONCLUSION

This study presents a hybrid Retrieval-Augmented Generation system for naturopathy-based health guidance — to the best of our knowledge, among the first to combine section-aware metadata filtering, dualmode hybrid retrieval, cross-encoder reranking, and two-layer semantic disease detection within a single locally deployed pipeline. Over a 30-query evaluation spanning 30 distinct diseases, the proposed system achieved Precision@3 of 0.644, Hit@3 of 0.767, and MRR of 0.767, compared to baseline scores of 0.322, 0.433, and 0.433 — improvements of 100%, 77%, and 77% respectively — with all improvements statistically significant at $p < 0.001$ (Wilcoxon signed-rank test).

Four architectural decisions drove these results. First, the 729 section-tagged chunks from The Encyclopedia of Natural Medicine provided a verified, disease-specific foundation. Second, the alias dictionary combined with SentenceTransformer-based semantic matching correctly handled informal query language — the cholesterol case (0.00 to 0.67) and the eczema case demonstrate this concretely. Third, hybrid ChromaDB and BM25 retrieval ensured coverage that neither strategy alone achieved. Fourth, cross-encoder reranking refined the merged candidate pool, with the ablation study confirming it as the single largest contributor to MRR improvement (0.45 to 0.90 in 10-query set). The full pipeline runs on a consumer laptop with a 4 GB GPU and 16 GB RAM without cloud API dependency — directly addressing the offline deployability gap identified in recent systematic reviews of healthcare AI [12].

Immediate future work priorities are: expanding the knowledge base beyond 42 diseases; depositing the dataset in a public repository (Zenodo); systematic threshold and chunk-size sensitivity analysis; hallucination evaluation using automated

frameworks such as RAGAS or TruLens; expert-annotated ground truth construction; and formal safety validation before any public or clinical deployment.

Declarations

Ethical approval

Not applicable

CRediT authorship contribution statement

Sakthivel K: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, and writing, editing, and reviewing of Original draft.

G.Rathi: Super-vision, Pooling of Resources, review and editing of the draft.

Shujithram K G: Software implementation, review and editing.

Thulasirajan S: Software implementation, review and editing.

Funding

Not applicable

Declaration of Competing Interests

The authors have affirmed their commitment to transparency by stating that they have no conflict of interest related to this article's research work, publication, and writing.

Consent for publication

Not applicable.

Availability of data and materials

Data are available upon reasonable request from the corresponding author.

Acknowledgment

The team is sincerely thankful to the Sri Ramakrishna Engineering College for their unwavering support. . Our heartfelt thanks also go to R K Nature Cure Hospital in Coimbatore, India for their consistent support and invaluable guidance throughout our project.

REFERENCES

[1] Bairy S, Rao MR, Edla SR, Manthena SR, Tatavarti NVGD. Effect of an integrated naturopathy and yoga program on long-term glycemic control in type 2 diabetes mellitus patients: A prospective cohort study. International Journal of Yoga. 2020;13(1):42-49. https://doi.org/10.4103/ijoy.IJOY_70_19

[2] Sharma S, Gupta A. Artificial intelligence in Ayurveda: Opportunities and challenges in traditional healthcare systems. Journal of Healthcare Engineering. 2022;2022:1-10. https://doi.org/10.1155/2022/4361236

[3] Suhardi, Rachman A, Nugroho MA. Development of an AI chatbot for herbal plant education using large language models. International Journal of Intelligent Systems and Applications. 2023;14(3):45-53. https://doi.org/10.5815/ijisa.2023.03.05

[4] Li Y, Zhang H, Wang J. Fine-tuning large language models for medical question answering with patient-doctor conversations. Proceedings of the IEEE International Conference on Artificial Intelligence in Medicine. 2023:112-118. https://doi.org/10.1109/AIME58045.2023.10223702

[5] Jurafsky D, Martin JH. Speech and Language Processing. 3rd ed. Stanford University; 2023. Available from: https://web.stanford.edu/~jurafsky/slp3/

[6] Green J. Phytotherapy in naturopathic practice: Evidence-based applications and outcomes. Journal of Traditional Medicine and Clinical Naturopathy. 2024;13(4):450. https://doi.org/10.37532/jtmcn.2024.13(4).450

[7] Li Y, Li S, Li Y. ChatDoctor: A medical chatbot based on large language models using retrieval-augmented generation. arXiv preprint arXiv:2303.14070. 2023. https://doi.org/10.48550/arXiv.2303.14070

[8] Vera J, Palaoag T. MedPlantBot: A medicinal plant-based healthcare chatbot using natural language processing. International Journal of Advanced Computer Science and Applications. 2021;12(6):450-457. https://doi.org/10.14569/IJACSA.2021.0120654

[9] Jansen M, Al-Farsi A, Petrovic L. Herbal remedies in diabetes management: Integrating Ayurveda with conventional treatments. Journal of Ayurveda and Naturopathy. 2026;3(1):14-17. https://doi.org/10.36648/ayurveda.3.1.014

[10] Somisetty KG, Shetty GB, Sujatha KJ, Shetty P. Naturopathy and yoga-based interventions modify risk factors of peripheral neuropathy and improve nerve conduction among type-2 diabetes mellitus patients. Journal of Ayurveda and Integrative Medicine. 2025;16. https://doi.org/10.1016/j.jaim.2024.100975

[11] Murray MT, Pizzorno JE. The Encyclopedia of Natural Medicine. 3rd ed. New York: Atria Books; 2012. Available from: https://www.simonandschuster.com/books/The-Encyclopedia-of-Natural-Medicine/MichaelT-Murray/9781451663006

[12] Huynh AL, Roy TJ, Jackson KN, Lee AG, Liaw W, Hossain MM. Applications of artificial intelligence-based conversational agents in healthcare: A systematic umbrella review. International Journal of Medical Informatics. 2026;207:106204. https://doi.org/10.1016/j.ijmedinf.2025.106204

[13] Li H, Zhang R, Lee YC, Sully R, Liu Y. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. npj Digital Medicine. 2023;6:236. https://doi.org/10.1038/s41746023-00979-5

[14] Pothuri VRV. Natural language processing and conversational AI. International Research Journal of Modernization in Engineering Technology and Science. 2024;6(9). https://doi.org/10.56726/IRJMETS61417

[15] Shaumya S, Kumar R. Artificial intelligence in Ayurveda: A systematic review (2020-2025). Journal of Ayurveda and Naturopathy. 2025;2(2):05-19. https://doi.org/10.33545/ayurveda.2025.v2.i2.A.24

[16] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-180. https://doi.org/10.1038/s41586-023-06291-2

[17] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617. 2023. https://doi.org/10.48550/arXiv.2305.09617

[18] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics. 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac409

[19] Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178. 2024. https://doi.org/10.48550/arXiv.2402.13178