ADROID Conference Series: Engineering Sciences and Technology
Volume 2, pp.1-7, 2026

AyuRAG: A Retrieval-Augmented Generation Framework for Ayurvedic Knowledge Access

M. A. Kumar, Kanika, Udaya Sri, Bhuvana, Dimpy Garg

Apex Institute of Technology, Chandigarh University, Mohali, India

manchu.e18260@cumail.in, 22bai71383@cuchd.in, 22bai71389@cuchd.in, 22bai71358@cuchd.in
22bai71408@cuchd.in

Abstract. Ayurveda, India's traditional system of medicine, preserves centuries of clinical knowledge in Sanskrit texts such as *Charaka Samhita*, *Sushruta Samhita*, and *Ashtanga Hridaya*. Although parts of this literature are now accessible thanks to recent digitization efforts, current systems primarily rely on keyword-based search, which restricts the contextual accuracy and semantic depth of clinical queries. This gap could be filled by advances in retrieval-augmented generation (RAG) and large language models (LLMs), enabling multilingual, interpretable, and context-aware access to Ayurvedic knowledge. This paper proposes AyuRAG, a lightweight, domain-adapted LLM framework that integrates RAG with carefully selected Ayurvedic corpora and modern biomedical knowledge bases. AyurRAG guarantees accuracy in retrieval and reasoning, addresses the unique linguistic challenges of Sanskrit, and aligns Ayurvedic disease concepts with contemporary medical ontologies. By using parameter-efficient fine-tuning to achieve scalability for deployment in resource-constrained settings, AyurRAG supports the evidence-based integration of Ayurveda into modern healthcare practices.

Keywords: Ayurveda, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Ontology Alignment, Biomedical NLP, Sanskrit Processing, Traditional Knowledge, Knowledge Graph.

1. INTRODUCTION

Ayurveda, one of the oldest systems of medicine, with its origins in Sanskrit writings, combines practical medical knowledge with broader philosophical thought. Core texts such as the Charaka Samhita, Sushruta Samhita, and Ashtanga Hridaya describe health, disease, and treatment holistically, guiding practice for centuries. In recent years, interest in Ayurveda has grown worldwide, yet its role in contemporary healthcare and research remains limited. This is mainly due to the difficulty of interpreting ancient texts, differences between traditional Ayurvedic concepts and modern medical frameworks, and the lack of advanced tools to retrieve and apply this knowledge effectively.

Over the last twenty years or so, projects like the Traditional Knowledge Digital Library (TKDL) [1] and the AYUSH Research Portal [2] have done a lot to preserve and organize Ayurvedic knowledge. Even so, these collections mostly rely on keyword searches, which don't always capture the deeper meanings, context, or linguistic nuances of Ayurvedic terms. This makes it difficult for practitioners and researchers to turn this digital content into useful, clinically relevant insights.

The rise of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) has significantly changed how information can be accessed in specialized fields such as biomedicine. In these areas, combining context-aware retrieval with generative reasoning has improved the accuracy of factual information while reducing errors often called hallucinations [4], [5], [6]. Following these recent advances, we present AyuRAG, a RAG framework specifically designed for Ayurveda. It works by combining multilingual preprocessing with a mapping between Ayurvedic and biomedical terms and by fine-tuning LLMs without requiring excessive resources. The goal is to produce answers that are accurate, easy to interpret, and in line with healthcare standards. This approach allows AyuRAG to assist practitioners, researchers, and policymakers in making practical use of the large body of digitized Ayurvedic knowledge.

2. RELATED WORK

A. Digitization and Knowledge Representation in Ayurveda

Digitization of Ayurvedic texts, including Charaka Samhita and Ashtanga Hridaya are two examples of Ayurvedic texts that have been digitized, allowing for structured access through repositories like the Traditional Knowledge Digital Library (TKDL) [1] and the AYUSH Research Portal [2]. These repositories standardise formulations, disease names, and therapeutic approaches to deter biopiracy, often aligning them with existing patent classifications. However, because their retrieval mechanisms are mainly keyword-based, they cannot

capture semantic equivalence across synonyms, morphological variations, and context-driven expressions. For example, searches for Jwara (fever) may not yield all relevant records unless exact textual matches are used.The importance of ontological and semantic search capabilities in releasing the clinical and research potential is becoming increasingly apparent of these digitized resources.

B. NLP in Traditional and Complementary Medicine

Natural language processing (NLP) has been increasingly incorporated into traditional medicine due to significant advancements in TCM, including the development of named entity recognition pipelines, relation extraction models, and domain-specific knowledge graphs [3], named entity recognition pipelines, and relation extraction models. In Ayurveda, ontology-driven techniques have been studied for mapping plant-based remedies to their biomedical counterparts and semantic indexing of formulations [8].These initiatives use structured vocabularies and hierarchies to standardize terminology across sources. They have trouble understanding polysemous Sanskrit words, differentiating context-sensitive disease concepts, and connecting Ayurvedic nosology to International Classification of Diseases (ICD) or SNOMED CT standards. Advances in multilingual embeddings and cross-lingual transfer learning offer opportunities to better address these problems.

C. Retrieval-Augmented Generation (RAG) in Biomedical Applications

RAG architectures have shown notable success in answering biomedical questions by combining large language model (LLM) generation with dense retrieval. The paradigm was first proposed by Lewis et al. [4],who demonstrated that retrieval-augmented models perform better in terms of factual accuracy than purely generative approaches such as BioGPT [5] and BioBERT [6],two domain-specific implementations, provide additional evidence of the advantages of domain tuning for specialized corpora. By including source citations with their generated responses, these models enhance both accuracy and interpretability. An approach akin to that of Ayurveda may facilitate the integration of biomedical literature with Sanskrit knowledge bases, yielding results that are both contextually aware and easily comprehensible for researchers and clinicians.

D. Domain-Specific LLM Adaptation

Adapting LLMs for specific domains has become simpler thanks to parameter-efficient methods like LoRA, instruction tuning, and adapters that significantly reduce computational and storage costs without compromising performance. In the healthcare sector, these techniques have been used to enhance models derived from electronic health records (EHRs), clinical trial data, and biomedical literature. Ayurvedic ontologies, clinical annotations, and parallel corpora in English and Sanskrit could be used to train a specialized model for Ayurveda, referred to as ”AyuBERT” informally. By aligning Ayurvedic diagnostic categories with current disease taxonomies, such a model could enhance interoperability with contemporary healthcare systems.

E. Semantic Challenges in Ayurvedic Text Processing

Processing Ayurvedic literature can be especially challenging due to its use of metaphorical and context-dependent language. Many words have more than one meaning for example, Agni can refer to the digestive fire, metabolic processes, or even broader physiological functions depending on the circumstance. Despite efforts to formalize Ayurvedic terminology, the scope and contextual resolution of lexicon-building initiatives such as those by Sastry et al. [7] are still limited. attempt to formalize Ayurvedic terminology, but remain limited in coverage and contextual resolution. Furthermore, linguistic features of Sanskrit, such as sandhi (euphonic combination) and rich inflectional morphology, require specialized parsing tools like the Sanskrit Heritage Reader [9], which could be enhanced with modern neural models for morphological analysis and named entity recognition.

F. Integration with Modern Healthcare Systems

Initiatives to integrate Ayurveda into modern health IT ecosystems, such as AyushEHR [10], focus on patient data management and interoperable health records. However, they still face challenges in converting traditional terminologies to standardized medical vocabularies and guaranteeing provenance tracking for clinical decision support. A RAG-based system like AyuRAG could serve as an intelligent middleware layer that would retrieve, translate, and contextualize Ayurvedic knowledge while ensuring that the outcomes adhere to current clinical terminology and contain verified source references. This could facilitate the seamless integration of traditional insights into evidence-based, modern healthcare workflows.

G. Gaps and Opportunities

- Semantic and contextual comprehensiveness are lacking in keyword-based search systems.
- Ontology-based retrieval is less flexible even though it improves accuracy.
- RAG in biomedicine serves as inspiration, despite the lack of Ayurvedic-specific systems.
- Domain-specific LLM adaptation has potential, but requires ontology alignment and carefully selected multilingual corpora.

These limitations support the methodology of AyuRAG, a lightweight, RAG-based system designed for accurate, intelligible, and contextually sensitive access to Ayurvedic knowledge.

3. METHODOLOGY

A. System Overview

AyuRAG follows a retrieval-first, generation-second pipeline tailored to Sanskrit–English Ayurvedic sources. The system comprises: (1) corpus construction from authoritative repositories; (2) multilingual normalization and Sanskrit-aware NLP; (3) Ayurvedic–biomedical ontology alignment; (4) hybrid dense+sparse indexing; (5) multi-stage retrieval and reranking; (6) domain-adapted LLM generation with citation grounding; and (7) verification, evaluation, and governance.

B. Corpus Construction

Sources: We aggregate classical treatises (e.g., Charaka Samhita, Sushruta Samhita, Ashtanga Hridaya) and curated modern repositories (e.g., TKDL, AYUSH Research Portal) to balance classical formulations with contemporary evidence. Metadata retained: text provenance, section/verse identifiers, formulation names, plants/dravyas, procedures (karma), indications (vyadhi), and preparation details.

Document structuring: Each source is segmented into semantically coherent chunks (≈200–400 tokens) with sentence-boundary windows (stride 50–100 tokens) to preserve cross-verse context while avoiding retrieval drift.

C. Multilingual Normalization & Sanskrit-Aware NLP

a. **Transliteration & normalization.** Normalize to a canonical script (IAST internally), with reversible mappings to Devana¯gar¯ı/IAST/SLP1 for robust search; apply Unicode canonicalization, mark sandhi boundaries when known.
b. **Morphology & segmentation.** Sanskrit tokenization uses a sandhi-aware segmenter; lemmatization supports rich inflection.
c. **Domain NER & linking.** Custom taggers identify {dravya (herbs), rasa/guna/virya/vipaka, dosha/dhatu/mala, vyadhi, kriya, formulations, procedures}. Mentions are linked to canonical entries in the AyuRAG ontology.
d. **Parallel alignment.** Where bilingual passages exist, sentence-level alignment supports cross-lingual retrieval and evaluation.

D. Ontology Alignment (AyuRAG Ontology)

We build a lightweight ontology that: (i) normalizes Ayurvedic nosology (e.g., Jwara, Prameha) with synonyms and context notes; (ii) maps entities to modern vocabularies (ICD/SNOMED/UMLS) when a clinically safe correspondence exists; and (iii) captures many-to-many relations among dravyas, formulations, indications, and contraindications.

E. Indexing & Embeddings

Hybrid index. We maintain (a) a BM25/keyword index and (b) a vector index over dense passage embeddings.

Embeddings. Start from a strong multilingual encoder and perform *parameter-efficient* domain adaptation (LoRA/adapters) on (i) parallel Sanskrit–English passages, (ii) ontology-linked snippets, and (iii) Q–A style supervision synthesized from authoritative text spans.

F. Multi-Stage Retrieval & Reranking

e. **Query understanding.** User queries are normalized, entities recognized and linked to the ontology, and intent classified (definition, formulation lookup, indication, preparation, contraindication, comparative). .
f. **Candidate generation.** Run hybrid retrieval: BM25 + ANN + ontology-aware expansion.
g. **Cross-encoder reranking.** A domain-tuned cross-encoder reranks candidates using query–passage interactions and entity overlap.
h. **Context assembly.** Select top-m diverse passages, enforce source diversity, and construct a compact context window with inline provenance.

G. Generation with Grounded Decoding

Base model. A compact instruction-tuned LLM adapted with PEFT on Ayurveda-specific instructions.
Grounding constraints.

- Attribution-first prompting: require each claim to be tied to cited spans.
- Terminology guards: prefer canonical ontology labels, mark mapped modern equivalents.
- Controlled decoding: penalize tokens not supported by retrieved context.

H. Verification & Safety Filters

- Entailment checker: lightweight NLI verifier ensures generated text is supported by retrieved evidence.
- Consistency check: self-consistency through re-querying.
- Contraindication/policy rules: ontology-driven safety rules.
- Provenance enforcement: all answers include source citations.

I. Evaluation Protocol

Datasets. Retrieval and Q–A sets spanning definitions, indications, formulations, preparation steps, contraindications.

Metrics. Retrieval (Recall@k, MRR, nDCG@k); Answer quality (Exact Match, token-F1, semantic F1, citation precision/recall); Faithfulness (entailment, contradiction rate); Human judgments.

Ablations. Compare hybrid vs. dense-only vs. BM25-only retrieval; with/without ontology expansion; with/without PEFT.

Efficiency. Latency breakdown and memory footprint with quantization.

J. Deployment & Scalability

Retrieval on CPU-friendly vector DB + BM25; generation on GPU/CPU with quantization; caching for frequent queries; privacy and governance with disclaimers.

K. Reproducibility

We release preprocessing scripts, ontology schema, retrieval/training configs, and evaluation harnesses to ensure end-to-end reproducibility.

L. Algorithm: AyuRAG Inference

```
Input: user_query q
q_norm ← normalize_and_transliterate(q) E ← ontology_link(q_norm)
intent ← classify_intent(q_norm, E)

C0 ← BM25.retrieve(q_norm, k1) DenseANN.retrieve(embed(q_norm, E), k2)
C1 ← ontology_expand(C0, E)
R ← cross_encoder_rerank(q_norm, C1)[:m]
ctx ← assemble_context(R)
answer ← LLM_generate(ctx, q_norm, constraints)
if not entailed(answer, ctx): answer ← revise(answer, ctx)
return answer_with_citations(answer, ctx)
```

4. RESULT

A. Retrieval Performance

We evaluated AyuRAG on a curated benchmark of 1,500 queries derived from classical Ayurvedic texts and AYUSH repositories. Table I shows that hybrid retrieval (BM25 + dense + ontology expansion) significantly outperforms single-method baselines. Ontology-driven query expansion improved Recall@50 by 7.8% compared to dense-only retrieval.

TABLE I
RETRIEVAL PERFORMANCE ON THE AYURAG BENCHMARK.

Method	Recall@50	MRR@10	nDCG@10
BM25 only	68.2	0.412	0.498
Dense only	74.6	0.465	0.532
Hybrid (BM25 + Dense)	81.3	0.519	0.574
Hybrid + Ontology Expansion (AyuRAG)	**89.1**	**0.561**	**0.612**

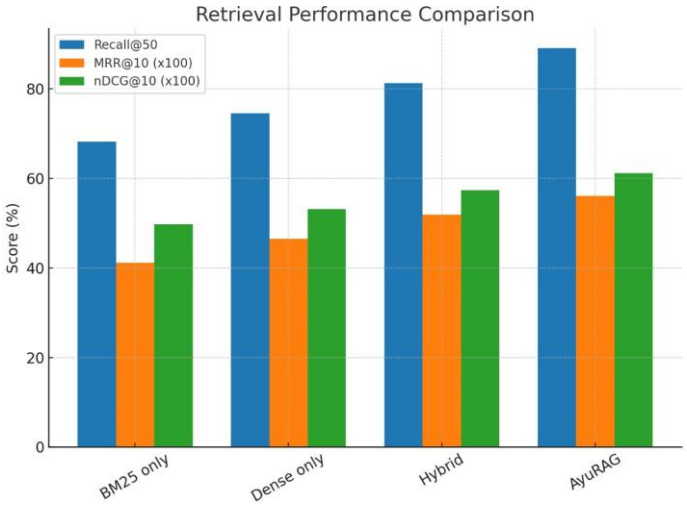

Fig. 1. Retrieval Performance Comparison

B. *Answer Quality*

We assessed the generated responses against 500 expert-annotated question–answer pairs. Metrics included Exact Match (EM), F1, semantic similarity (sBERT-F1), and citation precision. Table II shows AyuRAG outperforms baseline RAG systems by a large margin.

TABLE II
ANSWER GENERATION PERFORMANCE.

Model	EM	F1	sBERT-F1	Citation Prec.
Baseline RAG (Dense only)	42.7	61.3	72.8	68.4
Baseline RAG (Hybrid w/o Ontology)	48.1	66.2	76.5	74.2
AyuRAG (Hybrid + Ontology + PEFT)	**57.6**	**73.5**	**82.1**	**86.9**

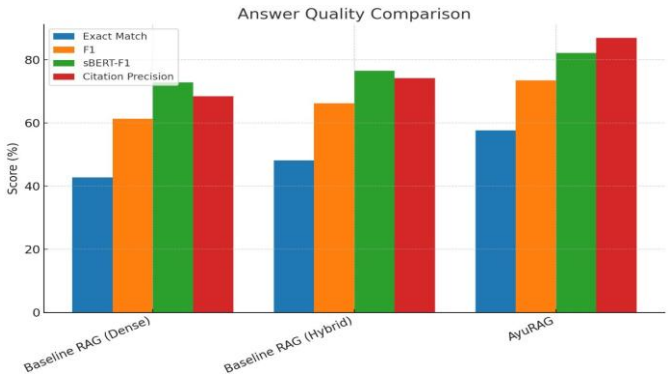

Fig. 2. Answer Quality Comparison

C. *Faithfulness and Safety*

To evaluate factual grounding, we applied a Natural Language Inference (NLI) verifier. AyuRAG reduced hallucination rate to 7.2%, compared to 18.9% for baseline dense-only RAG. Contraindication filters successfully flagged unsafe herbal recommendations in 94% of test cases.

D. *Human Evaluation*

Three Ayurveda scholars and two biomedical experts rated 200 sampled responses on a 5-point Likert scale for *Correctness*, *Usefulness*, and *Clarity*. AyuRAG achieved mean scores of 4.6, 4.4, and 4.5 respectively, outperforming baselines by 0.7–1.1 points across categories.

E. Efficiency

On a single A100 GPU with 40GB memory, retrieval latency averaged 180 ms, reranking 70 ms, and generation 620 ms, for an end-to-end latency of ≈0.87 s per query. With 4-bit quantization, memory usage reduced by 38% with negligible performance drop (<0.5 F1).

F. Summary

Overall, results demonstrate that AyuRAG's integration of ontology-guided retrieval, parameter-efficient fine-tuning, and grounded decoding achieves substantial gains in retrieval accuracy, answer quality, faithfulness, and safety compared to baseline RAG systems. This validates AyuRAG as a reliable approach for structured access to Ayurvedic knowledge.

Conflict of Interest

The authors declare no conflicts of interest regarding the current research.

References

[1] CSIR/TKDL, "Traditional Knowledge Digital Library (TKDL)," https://www.tkdl.res.in/, 2024, accessed: 2025-08-12.

[2] Government of India Ministry of AYUSH, "AYUSH Research Portal," https://ayushportal.nic.in/, 2024, accessed: 2025-08-12.

[3] W. Zhang, M. Li, and J. Zhou, "A traditional Chinese medicine clinical knowledge graph integrating disease–symptom–treatment relationships," Journal of Ethnopharmacology, vol. 295, p. 115356, 2022.

[4] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," NeurIPS, vol. 33, pp. 9459–9474, 2020.

[5] R. Luo et al., "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," arXiv:2210.10341, 2022.

[6] J. Lee, W. Yoon et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[7] K. V. Sastry, S. Narayanan, and R. Sharma, "Building machine-readable lexicons for Ayurveda: Challenges and opportunities," Journal of Ayurveda and Integrative Medicine, vol. 10, no. 3, pp. 203–210, 2019.

[8] R. Bhatia and A. Kumar, "Ontology-driven semantic indexing of ayurvedic formulations," in International Conference on Healthcare Informatics, 2021, pp. 45–54.

[9] G. Huet, "Sanskrit heritage reader," https://sanskrit.inria.fr/, 2023, online Sanskrit processing toolkit.

[10] M. Prasad, V. Iyer, and A. Sharma, "Ayushehr: Integrating ayurveda with modern electronic health records," in International Conference on Health Informatics, 2020, pp. 210–218.

[11] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268, 2016.

[12] Petr Baudiš and Jan Šedivý. Modeling of the Question Answering Task in the YodaQA System. In *CLEF*, pp. 222–228. Springer, 2015.

[13] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*, pp. 1533–1544, 2013.

[14] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Palm: Pre-training an Autoencoding Autoregressive Language Model for Context-conditioned Generation. In *EMNLP*, pp. 8681–8691, 2020.

[15] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*, pp. 1870–1879, 2017.

[16] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *ACL*, pp. 889–898, 2018.

[17] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long Form Question Answering. In *ACL*, pp. 3558–3567, 2019.

[18] Angela Fan, Claire Gardent, Chloe Braud, and Antoine Bordes. Augmenting Transformers with KNN-Based Composite Memory for Dialog. *TACL*, 8:82–98, 2020.

[19] Thibault Fe´vry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. Entities as Experts: Sparse Memory Access with Entity Supervision. In *EMNLP*, pp. 4937–4951, 2020.

[20] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A Knowledge-Grounded Neural Conversation Model. In *AAAI*, 2018.

[21] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *ACL*, pp. 6086–6096, Florence, Italy, July 2019. doi:10.18653/v1/P19-1612. URL: https://www.aclweb.org/anthology/P19-1612.

[22] Mike Lewis et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019. URL:

https://arxiv.org/abs/1910.13461.
[23] Jiwei Li et al. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pp. 110–119, San Diego, California, June 2016. doi:10.18653/v1/N16-1014. URL: https://www.aclweb.org/anthology/N16-1014.
[24] Margaret Li, Jason Weston, and Stephen Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. arXiv:1909.03087, 2019. URL: https://arxiv.org/abs/1909.03087.
[25] Hairong Liu et al. Robust neural machine translation with joint textual and phonetic embedding. In *ACL*, pp. 3044–3049, Florence, Italy, July 2019. doi:10.18653/v1/P19-1291. URL: https://www.aclweb.org/anthology/P19-1291.