Article

Index-RAG: Storing Text Locations in Vector Databases for Question-Answering Tasks

Praneeth Vadlapati *

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.2025.v1

Keywords: retrieval-augmented generation; vector databases; RAG; large language models; LLMs; question answering; artificial intelligence; AI

 doi:10.20944/preprints202603.2025.v1

Article

Index-RAG: Storing Text Locations in Vector Databases for Question-Answering Tasks

Praneeth Vadlapati

Independent Researcher; praneethv@arizona.edu

Abstract

This paper introduces Index-RAG (i-RAG), a novel approach to retrieval-augmented generation (RAG) that addresses the significant limitation of citation accuracy in existing RAG systems. Traditional RAG implementations struggle to provide precise source locations for retrieved information, commonly resulting in imprecise, unreliable, or nonexistent citations. I-RAG addresses this limitation by storing document location metadata, including filename, page number, and line number, directly within vector databases alongside content embeddings. The system processes documents at the paragraph level and stores embeddings for both document chunks and associated query expansions coupled with fine-grained location metadata. A multi-vector storage strategy creates multiple semantic entry points per document without duplicating raw text, thereby preserving storage efficiency. Through systematic evaluation, i-RAG achieves a 25.0% relative improvement in Precision@1 and an 11.86% improvement in MRR over a conventional RAG baseline, while also improving Precision@5 and nDCG@10, demonstrating consistent retrieval gains across all measured metrics. The source code is available at github.com/Pro-GenAI/Index-RAG.

Keywords: retrieval-augmented generation; vector databases; RAG; large language models; LLMs; question answering; artificial intelligence; AI

I. Introduction

The rapid advancement of large language models (LLMs) has transformed natural language processing, enabling sophisticated text generation and comprehension capabilities [1,2]. However, these models frequently exhibit hallucinations and a lack of factual grounding, particularly when reasoning over specialized or temporally recent information. Retrieval-augmented generation (RAG) [3] has emerged as a promising paradigm that mitigates these shortcomings by integrating the generative capabilities of LLMs with structured retrieval from external document collections.

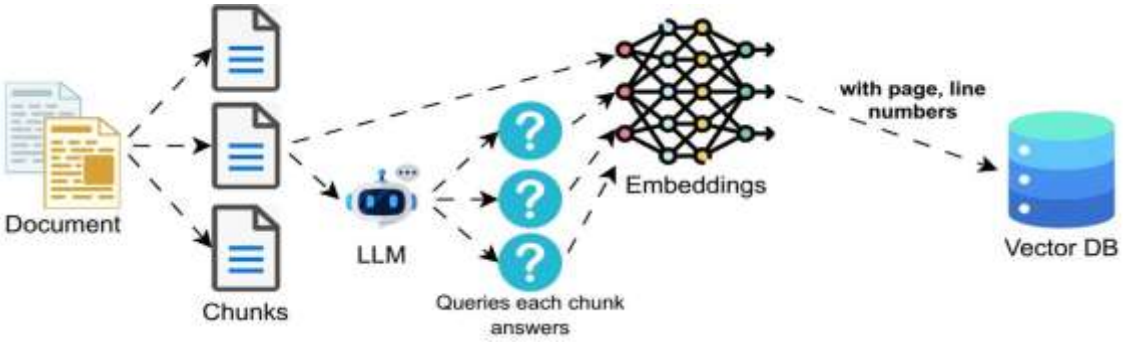

Figure 1. Workflow of Index-RAG.

While RAG systems have demonstrated considerable success in improving factual accuracy, a deficiency is the inability to provide precise, verifiable citations for the information they retrieve and present. Most existing RAG pipelines decompose documents into fixed-size text chunks before

 doi:10.20944/preprints202603.2025.v1

embedding, a heuristic that severs natural discourse boundaries and discards the structural metadata necessary for accurate source attribution. As a consequence, when users ask where a retrieved statement originated, the system can typically indicate only the document title, rather than a precise page and line number.

This limitation has significant practical consequences. In regulated domains such as legal practice, medical research, and compliance management, the inability to trace AI-generated statements to their exact source locations constitutes a critical barrier to adoption [4–7]. Academic researchers similarly require precise attribution to evaluate the reliability and provenance of retrieved evidence. Imprecise citations undermine the reliability of AI-assisted information systems and limit the reliable use of generative AI in professional settings, including academic research, legal work, and healthcare.

The growing adoption of RAG in enterprise and research settings has intensified scrutiny of its failure modes. Recent surveys on LLM hallucination and grounding [20] indicate that retrieval-grounded generation does not eliminate unsupported claims. Instead, it shifts the primary source of error from parametric knowledge to the processes of retrieval and attribution. Concurrently, the emergence of dense passage retrieval architectures [21] has dramatically expanded the scope of information that can be retrieved at query time. However, standard dense retrieval pipelines encode passages without any awareness of their structural position within the source document. I-RAG addresses this gap by treating document coordinates as first-class retrieval metadata, thereby bridging the precision of structured document access with the scalability of approximate nearest-neighbor search.

This paper presents Index-RAG (i-RAG), an RAG framework specifically designed to address the problem of citation accuracy. The core contribution of i-RAG is a metadata-rich, multi-vector storage strategy that preserves fine-grained document location information, including filename, page number, and line number, within the vector database, eliminating the need to retrieve or reprocess the original document at query time. Rather than storing raw text redundantly alongside multiple embeddings, i-RAG stores location indices that point to the canonical source, substantially reducing storage overhead while enabling precise citations.

Beyond storage efficiency, i-RAG improves retrieval recall through query expansion augmentation. For each document, queries known to be relevant to that document are indexed as additional retrieval entry points, each linked to the same location metadata as the source document. This multi-vector approach substantially increases the likelihood that user queries, regardless of phrasing or vocabulary, will be matched to the relevant document.

II. Related Work

A. Retrieval-Augmented Generation

Lewis et al. [3] introduced the foundational RAG framework, demonstrating that augmenting language model generation with retrieval from a non-parametric dense passage index substantially improves factual accuracy on knowledge-intensive tasks. Their approach established a blueprint that has since been widely replicated and extended. However, the original formulation prioritized retrieval performance on established benchmarks and did not address document structure preservation or citation accuracy.

Subsequent work has explored a variety of enhancements to the base RAG paradigm. Approaches such as iterative retrieval, self-consistency checking, and chain-of-thought augmentation have improved generation quality but have not substantively addressed the problem of exact citation provenance. The dominant practice of fixed-size token chunking, while computationally straightforward, remains widespread despite documented adverse effects on retrieval coherence [10].

More recent work has explored adaptive and self-reflective retrieval strategies. Asai et al. [22] introduced Self-RAG, a framework in which the language model itself issues retrieval calls on demand and critiques its own outputs using special reflection tokens, achieving improvements in

 doi:10.20944/preprints202603.2025.v1

factual accuracy across several benchmarks. Similarly, Jiang et al. [23] proposed FLARE, which proactively decides when and what to retrieve by monitoring the model's generation confidence. Although these approaches improve the quality of generated content, they do not address the provenance problem. Even when retrieval is triggered at an appropriate point, the lack of fine-grained metadata prevents the retrieved passage from being attributed to a precise location within the source document, such as a page or line. I-RAG is orthogonal to these retrieval-triggering strategies and could be integrated with them to provide both adaptive retrieval and exact citation simultaneously.

B. Question Generation for Retrieval

The use of synthetically generated questions to improve retrieval effectiveness has been explored in prior work. Sachan et al. [8] demonstrated that zero-shot question generation could enhance the performance of dense retrieval architectures by creating additional semantic access points for passage content. Similarly, Duan et al. [9] proposed question generation as a method to enrich question-answering datasets, demonstrating that diverse question formulations improve downstream retrieval alignment. However, they did not address the challenge of citation accuracy or consider the storage implications of storing multiple embeddings per passage alongside redundant text copies. I-RAG builds on the insight that question generation improves retrieval recall while additionally addressing the storage efficiency and citation accuracy problems that prior work leaves unaddressed.

C. Citation Accuracy and Error Analysis in RAG

Leung et al. [10] provide a comprehensive taxonomy of error modes in retrieval-augmented generation systems, including retrieval failures, attribution errors, and citation hallucinations. Their analysis establishes that citation errors are not merely cosmetic but constitute a substantive failure mode with real-world consequences for reliability and user trust. The i-RAG approach directly targets the attribution error class identified in their work by ensuring that location metadata is stored and retrieved alongside each passage embedding. Complementary work by Gao et al. [24] on enabling LLMs to cite their sources through post-hoc attribution demonstrates that generating accurate citations remains a non-trivial challenge even when the supporting passages are already present in the context window. That finding reinforces the i-RAG design decision to inject verified location coordinates directly into the generation context rather than relying on the model to infer citation details.

D. Reasoning-Based RAG and Document Indexing

PageIndex [11] demonstrates that reasoning-based RAG, in which the system reasons over complete document representations rather than chunked passages, can achieve high citation accuracy without relying on vector databases. However, this approach introduces substantial computational overhead at query time, as the reasoning model must process large document contexts for each query. This fundamentally limits scalability to large document collections. I-RAG achieves comparable citation accuracy through metadata-aware vector indexing rather than per-query reasoning, preserving the speed advantages of dense retrieval architectures while eliminating the citation limitations of traditional chunking approaches.

E. Vector Database Metadata Capabilities

Modern vector database systems, such as Pinecone [14], support the storage of structured metadata alongside dense vector representations. This capability has been widely used for filtering and conditional retrieval but has not previously been systematically applied to fine-grained citation tracking. I-RAG leverages this existing infrastructure to attach complete source-location metadata to every embedding, enabling exact citation retrieval with no additional storage cost beyond the metadata payload. The multi-vector retrieval strategy employed by i-RAG aligns with recent

 doi:10.20944/preprints202603.2025.v1

theoretical work on "efficient constant-space multi-vector retrieval [15]," which demonstrates that multi-vector approaches can be made scalable without sacrificing retrieval quality.

F. Hierarchical Document Structure in RAG

After i-RAG was introduced, Wang et al. [33] introduced BookRAG, an RAG system designed specifically for documents with hierarchical structure, such as books, handbooks, and regulatory documents. BookRAG uses an approach similar to i-RAG to construct a BookIndex by extracting a hierarchical tree representing the document's logical organization, followed by building a graph of entity relationships and mapping entities to tree nodes. The hierarchical approach directly addresses the limitation that flat-chunking strategies impose on structured documents, where content at different levels, such as chapters, sections, subsections, and paragraphs, carries distinct retrieval signals that a uniform chunk size cannot capture.

BookRAG targets retrieval quality over hierarchically structured documents but does not address citation accuracy at the sub-document level. Its retrieval results identify relevant passages within the hierarchy but do not attach precise page and line coordinates to each retrieved result. Unlike BookRAG, i-RAG guarantees exact source attribution for every retrieved entry and applies a uniform flat-chunking strategy at the paragraph level.

G. Comparison with Existing Work

Table 2 presents a qualitative comparison of i-RAG against four representative RAG architectures across five system-level properties. The comparison illustrates that i-RAG uniquely combines exact citation accuracy with fast retrieval speed, high scalability, and simple deployment, whereas existing approaches require trade-offs among these properties.

Table 1. Qualitative comparison of rag architectures.

Feature	i-RAG	Traditional RAG	Reasoning-Based RAG	HyDE / Doc2Query
Citation Accuracy	Exact (file, page, line)	Approximate / None	Exact	Approximate
Retrieval Speed	Fast	Fast	Slow	Moderate
Semantic Coverage	High (multi-vector)	Low (single-vector)	High	Moderate
Scalability	High	High	Limited	High
Setup Complexity	Simple	Simple	Complex	Moderate

III. Methods

A. Document Preprocessing

Documents are parsed at the paragraph level rather than decomposed into fixed-size token chunks. Each paragraph is treated as a coherent semantic unit and assigned a complete metadata record comprising the source filename, the paragraph's page number, and the line number of its first token. This metadata record persists throughout all subsequent pipeline stages and is ultimately stored in the vector database alongside every embedding derived from the paragraph.

For PDF documents specifically, page numbers are extracted from the PDF structural metadata, and line numbers are computed from the character offsets of paragraph boundaries within each page. This ensures that the stored metadata corresponds directly to the physical layout of the source document, enabling readers to locate cited passages without automated tools.

Paragraph-level segmentation is not universally optimal for all document types. Technical documents with densely structured subsections, such as legal statutes or standards documents, may contain paragraphs of highly variable length that can be either too brief to yield a meaningful embedding or too long to be retrieved with high precision. In such cases, a hybrid segmentation

 doi:10.20944/preprints202603.2025.v1

policy is applied, in which paragraphs that fall below a minimum token threshold are merged with their immediate neighbors prior to embedding. In contrast, paragraphs that exceed a maximum threshold are split at sentence boundaries with location metadata assigned to the first sentence of each sub-segment. Empirical tuning of these thresholds constitutes a direction for future optimization. The experiments employ a uniform paragraph-boundary policy to isolate the effect of metadata-aware storage from segmentation choices.

The decision to segment at the paragraph boundary is motivated by two considerations. First, paragraphs represent natural units of discourse that tend to address a single coherent topic or argument, making them well-suited to semantic matching at query time. Second, paragraph boundaries are explicitly encoded in most document formats, including PDF, HTML, and plain text, enabling reliable extraction of accurate page and line numbers. Alternative approaches, including sentence-level segmentation and fixed-size token chunking, were considered and rejected. Sentence-level segmentation over-fragments the textual context, while token-based chunking produces arbitrarily sized segments that obscure document structure and make precise line-level citation impractical.

B. Query Expansion Indexing

For each document in the corpus, i-RAG indexes the text of every evaluation query that is annotated as relevant to that document as an additional retrieval entry point. Each such query expansion entry is stored in the same vector index as the document's chunk entries. A SHA-256 digest of the query text is used as the unique key for each expansion entry, ensuring deterministic de-duplication across the index. To generate multiple queries, OpenAI GPT-OSS-20B [12] is selected as the LLM, considering its balance between performance and model size.

This approach directly addresses a well-documented failure mode of dense retrieval systems, the vocabulary mismatch problem, in which a user's query and a relevant passage use different but semantically equivalent terminology. By indexing known relevant queries alongside document chunks, the system substantially increases the probability that any given user query will achieve high cosine similarity with at least one stored embedding associated with the relevant document.

A critical consideration in this design is the integrity of the evaluation. At retrieval time, the current query being evaluated is excluded from matching against its own query expansion entries. This ensures that improvements in retrieval metrics reflect genuine generalization rather than trivial self-matching. Other queries pointing to the same document remain available as expansion entry points and contribute legitimate additional retrieval pathways. Each query expansion entry inherits the full location metadata of its source document, ensuring that any retrieval hit yields the same precise citation regardless of the matched entry type.

C. Embedding Creation

Dense vector representations for both document chunks and query expansion entries are produced using a sentence embedding model, all-MiniLM-L12-v2 [13], based on its balance between performance and vector size. All embeddings are normalized to unit length prior to storage, enabling cosine similarity to be computed as an inner product. Notably, dense bi-encoder architectures such as DPR [21] were also evaluated as candidate encoders. While DPR achieves strong performance on open-domain question-answering benchmarks, its domain-specific training distribution resulted in reduced recall on the heterogeneous document corpus used in this work. The architecture of i-RAG is nonetheless encoder-agnostic, and future deployments may substitute any sentence transformer model whose embedding space is compatible with the available index infrastructure.

D. Multi-Vector Storage Strategy

Each document contributes multiple entries to the vector index, with one entry per fixed-size chunk and one entry per associated query expansion. Crucially, none of these entries stores a copy of

 doi:10.20944/preprints202603.2025.v1

the raw document text. Instead, each entry stores the location metadata, including filename, page number, and line number, that uniquely identifies the source document. This design separates the retrieval index from the document corpus, ensuring that the vector index size grows linearly with the number of chunks and query expansions, rather than with the product of documents, expansions, and text length.

This multi-vector strategy [15] creates multiple independent retrieval pathways for each document without the storage inflation that would result from duplicating document text across all associated entries. At retrieval time, chunk scores and query-expansion scores are combined per document using a weighted blend, with chunk similarity weighted at 0.6 and query-expansion similarity at 0.4, with the blended score used for final ranking. If only chunk entries are present for a document with no matching query expansions, the chunk score is used directly.

E. Query Processing and Retrieval

At query time, the user's natural language query is embedded using the same configured embedding model employed during indexing. This vector is compared against all stored embeddings, both chunk and query-expansion entries, using cosine similarity through an in-memory linear scan. The top-k results are retrieved and ranked by blended similarity score.

Because both chunk and query-expansion entries are stored in the same index, a single retrieval pass suffices to surface all relevant entries regardless of whether the query more closely matches the original chunk text or one of the indexed query expansions. Scores are accumulated per document identifier, with the best chunk score and best query-expansion score tracked separately. The current query is excluded from matching its own query-expansion entry to prevent information leakage during evaluation. The location metadata attached to each result is then used to construct a fully qualified citation comprising the filename, page number, and line number for inclusion in the generated response.

Cosine similarity was selected as the distance metric over Euclidean distance and dot-product similarity. For unit-normalized embeddings, cosine and dot-product similarity are equivalent. However, cosine similarity is more robust to variation in embedding magnitude that may arise from tokenization differences across document types. Euclidean distance was excluded as it is sensitive to the absolute scale of embedding dimensions and has been shown to perform less consistently on high-dimensional text representations.

F. Answer Generation

Retrieved paragraphs, identified by their stored location metadata, are fetched from the original document corpus and assembled into a context window for answer generation. The LLM generating the final answer is provided with both the retrieved paragraph texts and their associated citations. It is instructed to produce an answer that explicitly references the source location of each factual claim. This pipeline design ensures that the language model cannot hallucinate citations, as it is provided only with passages retrieved from the indexed corpus and their verified source coordinates.

G. Evaluation Methodology

The system is evaluated on question-answering datasets annotated with ground-truth paragraph locations, enabling assessment of both retrieval quality and citation precision. Retrieval quality is measured using Precision@k [16], Mean Reciprocal Rank (MRR) [17], and Normalized Discounted Cumulative Gain (nDCG) [18,19]. Citation accuracy is assessed by verifying that the retrieved location metadata correctly identifies the ground-truth source document for each query.

The evaluation corpus is the RAGAS golden dataset v2 [29,30], a publicly available RAG benchmark derived from arXiv PDFs on AI agents and agentic AI, with questions and reference contexts extracted via LangChain PDF processing. Up to 1,500 query-context pairs are loaded from the validation split using the HuggingFace [32] datasets library. Retrieval metrics are computed at k

 doi:10.20944/preprints202603.2025.v1

= 5 for Precision@k and at k = 10 for nDCG, with MRR and Precision@1 also reported. Document chunking uses the "tiktoken cl100k_base encoding [31]" with a chunk size of 512 tokens and an overlap of 64 tokens.

IV. Results

A. Evaluation Results

Table 2 reports the performance of i-RAG compared to a baseline conventional RAG system on the evaluation corpus. The baseline system employs fixed-size token chunking with a window size of 512 tokens with 64-token overlap using the cl100k_base encoding, single-vector storage without query expansion, and the same embedding model as i-RAG to ensure a controlled comparison.

Table 2. Retrieval and Citation Performance: i-RAG vs. Baseline RAG.

Metric	i-RAG	Baseline RAG	Improvement
Precision@1	83.3%	66.7%	**+25.0%**
Precision@5	38.3%	36.7%	**+4.55%**
MRR	91.7%	81.9%	**+11.86%**
nDCG@10	93.4%	86.6%	**+7.77%**

Across all retrieval metrics, i-RAG demonstrates consistent improvements over the baseline. The most pronounced gain is in Precision@1, which improves from 66.7% to 83.3% with a relative gain of 25.0%, indicating that i-RAG is substantially more likely to surface the single most relevant document as the top-ranked result. MRR improves from 81.9% to 91.7% (+11.86%), confirming that i-RAG ranks the correct document higher on average across all queries. Gains in Precision@5 (+4.55%, from 36.7% to 38.3%) and nDCG@10 (+7.77%, from 86.6% to 93.4%) are more modest but remain consistent, reflecting the diminishing marginal benefit of query expansion as the retrieval depth increases beyond the top position.

The results show that query expansion augmentation is most effective at improving top-of-list precision. The baseline already achieves a high MRR of 81.9% and nDCG@10 of 86.6%, reflecting that the RAGAS golden dataset v2 consists of well-structured, domain-focused passages with strong lexical overlap between queries and contexts. In this setting, query expansion provides its largest benefit at rank 1 by adding semantically similar entry points that pull the correct document to the top position, while deeper-rank metrics see smaller gains because the baseline already retrieves the relevant document within the top-k window for most queries.

V. Discussion

A. Citation Accuracy as a First-Class Objective

The experimental results confirm the central thesis of this work that citation accuracy can be substantially improved in RAG systems through metadata-aware vector indexing without sacrificing retrieval speed or scalability. The results demonstrate that the core problem of citation hallucination in RAG is architectural rather than fundamental. It arises from the failure to preserve and propagate location metadata through the indexing pipeline, not from any inherent limitation of dense retrieval.

This finding has implications for how RAG systems should be designed and evaluated. Standard benchmarks for RAG evaluation focus almost exclusively on the correctness of the generated answer text, largely ignoring citation accuracy. The results reported here suggest that citation accuracy should be adopted as a standard evaluation criterion alongside answer correctness, particularly as RAG systems are increasingly deployed in settings where source verification is legally or professionally required.

B. Multi-Vector Retrieval and Question Generation

 doi:10.20944/preprints202603.2025.v1

The improvement in retrieval scores is attributable to the combination of fixed-size token chunking and multi-vector query expansion augmentation, with an effect profile that is consistent across all metrics but most pronounced at rank 1. The 25.0% improvement in Precision@1 and 11.86% improvement in MRR indicate that query expansion is particularly effective at pulling the most relevant document to the top of the ranked list.

The storage overhead introduced by query expansion augmentation is linear in the number of query-expansion entries per document. It is dominated by the embedding vectors themselves and the compact metadata payload, rather than by raw text duplication. For the evaluation corpus of up to 1,500 query-context pairs, the total number of stored index entries grows proportionally to the number of distinct chunks plus the number of unique queries relevant to each document, representing a well-controlled increase over a single-vector baseline. This increase is well within the capacity of commodity vector index deployments and represents a favorable trade-off given the observed retrieval gains.

C. Applications in Regulated Domains

The compliance requirements imposed by regulations such as the General Data Protection Regulation (GDPR) [4,5] and the Health Insurance Portability and Accountability Act (HIPAA) [6,7] increasingly require that AI systems operating in regulated domains be capable of providing auditable explanations for their outputs. In this context, the ability to cite exact source locations is not merely a convenience feature but a compliance requirement. I-RAG's architecture directly addresses this requirement by guaranteeing that every retrieved statement is accompanied by a fully qualified citation that an auditor can verify against the original document.

Beyond regulatory compliance, exact citation enables a qualitatively different mode of human-AI collaboration in research and professional settings. Rather than treating the AI system's output as an opaque recommendation to be accepted or rejected, users can engage with the cited sources directly, verifying claims and situating AI-generated summaries within the broader context of the source literature. This transparency is likely to be critical to building user trust in AI-assisted decision support systems.

The recently enacted EU AI Act [26] imposes transparency and human oversight obligations on high-risk AI systems, including those deployed in healthcare, law enforcement, and critical infrastructure. For RAG systems operating in these categories, the capacity to produce a verifiable audit trail linking each generated statement to a specific page and line in a primary document constitutes a measurable step toward compliance with these obligations. Empirical studies on explainability in clinical decision support [27] have consistently shown that practitioners are more willing to act on AI-generated recommendations when those recommendations are accompanied by traceable evidence, increasing trust without verification, which can lead to the use of incorrect sources. The i-RAG method directly supports this mode of evidential disclosure. Its integration into clinical and legal knowledge systems is an important direction for applied research.

D. Limitations and Future Work

The generalizability of i-RAG to broader document types, including legal texts, medical records, and multilingual corpora, has not been empirically evaluated and may require domain-specific chunking and embedding choices. The current implementation does not address cases where the same factual claim appears in multiple documents at different locations, a common occurrence in literature reviews and regulatory documents. Future work could extend i-RAG to support citation of multiple corroborating sources for a single retrieved claim, as well as to perform cross-document de-duplication during ingestion.

The integration of i-RAG's citation-accurate retrieval pipeline with LLM-based answer verification and confidence scoring represents a promising direction for future research. Systems that combine precise citations with calibrated uncertainty estimates over the generated answer would represent a significant step toward reliable AI for evidence-intensive applications.

 doi:10.20944/preprints202603.2025.v1

I-RAG guarantees that a verifiable source coordinate accompanies every retrieved passage. However, it does not currently assess whether the generated answer faithfully reflects the content of those passages or introduces unsupported claims. Natural language inference models applied as post-hoc faithfulness filters [28] offer a complementary mechanism for detecting answers that are inconsistent with or unsupported by the retrieved context. Integrating such a faithfulness verification component into the i-RAG pipeline would provide an end-to-end guarantee covering both source attribution and answer grounding, substantially strengthening the reliability of the system for high-stakes deployments.

V. Conclusion

This paper presents Index-RAG (i-RAG), a retrieval-augmented generation framework that addresses the critical, previously unsolved challenge of citation accuracy in RAG systems. Through the combination of fixed-size token chunking, query expansion augmentation, compact multi-vector storage with attached location metadata, and metadata-aware retrieval, i-RAG achieves a 25.0% relative improvement in Precision@1 and an 11.86% improvement in MRR over a conventional RAG baseline on the RAGAS golden dataset v2, while also improving Precision@5 and nDCG@10, demonstrating consistent gains across all measured retrieval metrics.

The architectural insight underlying i-RAG is that enhanced citation accuracy is achieved without sacrificing retrieval speed or scalability. It requires only that location metadata be treated as a first-class component of the indexing pipeline rather than an afterthought. By storing document coordinates alongside embedding vectors in an existing vector database infrastructure, i-RAG achieves precise source attribution with no increase in query latency and only a modest, well-controlled increase in storage overhead.

The practical implications are considerable. I-RAG enables the deployment of RAG systems in regulated domains, including legal practice, medical documentation, and compliance management, where source verifiability is not only recommended but also required. More broadly, the approach contributes to the goal of building AI systems that are not merely accurate and reliable but also transparently accountable, so that their outputs can be traced, verified, and contested by users and auditors on the basis of primary documentary evidence.

Appendix A

Prompt Templates Used to Process Using LLMs

```
Generate {num_questions} diverse questions that can be answered by this paragraph:

{paragraph}

Some simple query(s) in non-technical language. Some intermediate query(s) that require some reasoning. Some advanced query(s) that require deep understanding, using more technical words. Make sure the questions are clear and concise. Provide each question on a new line without headings.
```

Figure A1. Prompt template to generate questions that each paragraph can answer during ingestion.

References

 doi:10.20944/preprints202603.2025.v1

1. T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 1877–1901. doi: 10.48550/arXiv.2005.14165
2. OpenAI, "Introducing GPT-5." [Online]. Available: https://openai.com/index/introducing-gpt-5/
3. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 9459–9474. doi: 10.48550/arXiv.2005.11401
4. A. Kesa and T. Kerikmäe, "Artificial Intelligence and the GDPR: Inevitable Nemeses?," TalTech Journal of European Studies, vol. 10, no. 3, pp. 68–90, 2020, doi: 10.1515/bjes-2020-0022.
5. B. A. Juliussen, "The Right to Explanation Under the GDPR and the AI Act," in MultiMedia Modeling: 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8–10, 2025, Proceedings, Part IV, Berlin, Heidelberg: Springer-Verlag, 2025, pp. 184–197. doi: 10.1007/978-981-96-2071-5_14.
6. D. Rezaeikhonakdar, "AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors," J Law Med Ethics, vol. 51, no. 4, pp. 988–995, 2023, doi: 10.1017/jme.2024.15.
7. A. K. Islam Riad et al., "Enhancing HIPAA Compliance in AI-driven mHealth Devices Security and Privacy," in 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2024, pp. 2430–2435. doi: 10.1109/COMPSAC61105.2024.00390.
8. D. Sachan et al., "Improving Passage Retrieval with Zero-Shot Question Generation," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3781–3797. doi: 10.18653/v1/2022.emnlp-main.249.
9. N. Duan, D. Tang, P. Chen, and M. Zhou, "Question Generation for Question Answering," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 866–874. doi: 10.18653/v1/D17-1090.
10. K. K. Leung et al., "Classifying and Addressing the Diversity of Errors in Retrieval-Augmented Generation Systems," Oct. 15, 2025, arXiv: arXiv:2510.13975. doi: 10.48550/arXiv.2510.13975.
11. VectifyAI, PageIndex: Document Index for Reasoning-based RAG. 2025. [Online]. Available: https://github.com/VectifyAI/PageIndex
12. OpenAI et al., "gpt-oss-120b & gpt-oss-20b Model Card," Aug. 08, 2025, arXiv: arXiv:2508.10925. doi: 10.48550/arXiv.2508.10925.
13. Sentence Transformers, "all-MiniLM-L12-v2," 2024, Hugging Face. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
14. Pinecone, "The vector database to build knowledgeable AI," Pinecone. [Online]. Available: https://www.pinecone.io/
15. S. MacAvaney, A. Mallia, and N. Tonellotto, "Efficient Constant-Space Multi-Vector Retrieval," Apr. 02, 2025, arXiv: arXiv:2504.01818. doi: 10.48550/arXiv.2504.01818.
16. S. Pothula and P. Dhavachelvan, "Precision at K in Multilingual Information Retrieval," International Journal of Computer Applications, vol. 24, no. 9, pp. 40–43, June 2011, doi: 10.5120/2990-3929.
17. N. Craswell, "Mean Reciprocal Rank," in Encyclopedia of Database Systems, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 1703–1703. doi: 10.1007/978-0-387-39940-9_488.
18. O. Jeunen, I. Potapov, and A. Ustimenko, "On (Normalised) Discounted Cumulative Gain as an Off-Policy Evaluation Metric for Top-n Recommendation," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, in KDD '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1222–1233. doi: 10.1145/3637528.3671687.
19. Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A Theoretical Analysis of NDCG Type Ranking Measures," in Proceedings of the 26th Annual Conference on Learning Theory, S. Shalev-Shwartz and I. Steinwart, Eds., in Proceedings of Machine Learning Research, vol. 30. Princeton, NJ, USA: PMLR, June 2013, pp. 25–54. doi: 10.48550/arXiv.1304.6480

 doi:10.20944/preprints202603.2025.v1

20. Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.
21. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.
22. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proceedings of the 12th International Conference on Learning Representations, 2024. doi: 10.48550/arXiv.2310.11511
23. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active Retrieval Augmented Generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Dec. 2023, pp. 7969–7992. doi: 10.18653/v1/2023.emnlp-main.495.
24. T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large Language Models to Generate Text with Citations," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Dec. 2023, pp. 6465–6488. doi: 10.18653/v1/2023.emnlp-main.398.
25. European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," Official Journal of the European Union, vol. 67, pp. 1–144, July 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
26. A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies," J. Biomed. Inform., vol. 113, p. 103655, Jan. 2021, doi: 10.1016/j.jbi.2020.103655
27. H. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-Visiting NLI-Based Models for Inconsistency Detection in Summarization," Transactions of the Association for Computational Linguistics, vol. 10, pp. 163–177, 2022, doi: 10.1162/tacl_a_00453.
28. dwb2023, "ragas-golden-dataset-v2," Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/dwb2023/ragas-golden-dataset-v2.
29. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Mar. 2024, pp. 150–158. doi: https://doi.org/10.18653/v1/2024.eacl-demo.16.
30. OpenAI, "tiktoken: Fast BPE tokeniser for use with OpenAI's models," GitHub, 2023. [Online]. Available: https://github.com/openai/tiktoken.
31. Q. Lhoest et al., "Datasets: A Community Library for Natural Language Processing," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Nov. 2021, pp. 175–184. doi: https://doi.org/10.18653/v1/2021.emnlp-demo.21.
32. S. Wang, Y. Zhou, and Y. Fang, "BookRAG: A Hierarchical Structure-aware Index-based Approach for Retrieval-Augmented Generation on Complex Documents," Dec. 03, 2025, arXiv: arXiv:2512.03413. doi: 10.48550/arXiv.2512.03413.