Index-RAG: Storing Text Location in Vector Databases for QA tasks

Praneeth Vadlapati
Independent researcher
praneethv@arizona.edu
ORCID: 0009-0006-2592-2564

Abstract: This paper introduces Index-RAG (i-RAG), a novel approach to retrieval-augmented generation (RAG) that addresses the critical limitation of citation accuracy in existing RAG systems. Traditional RAG implementations struggle to provide precise source locations for retrieved information, often resulting in vague or inaccurate citations. I-RAG solves this problem by storing document location metadata directly within vector databases alongside content embeddings. The system processes documents at the paragraph level, generates multiple relevant questions for each paragraph using large language models, and stores embeddings for both the questions and paragraphs with precise location coordinates including filename, page number, and line number. Through comprehensive evaluation on question-answering benchmarks, i-RAG demonstrates superior citation accuracy while maintaining competitive retrieval performance. The approach represents a significant advancement in making RAG systems more trustworthy and suitable for applications requiring source verification, such as academic research, legal analysis, and compliance documentation.

The source code is available at github.com/Pro-GenAI/Index-RAG.

Keywords: Large Language Models, LLMs, Retrieval-Augmented Generation, Vector Databases, Citation Accuracy, Question Answering, Document Location Tracking, Artificial Intelligence, AI

I. Introduction

The rapid advancement of large language models (LLMs) has revolutionized natural language processing, enabling sophisticated text generation and comprehension capabilities [1], [2]. However, these models often suffer from hallucinations and lack of factual grounding, particularly when dealing with specialized or current information. Retrieval-augmented generation (RAG) [3] has emerged as a promising solution by combining the generative power of LLMs with external knowledge retrieval from document collections.

A. Disadvantages with Current Approaches

Despite significant progress in RAG systems, a critical challenge remains: the inability to provide accurate and precise citations for retrieved information. Traditional RAG implementations typically chunk documents into arbitrary text segments and store these chunks as vectors, losing the original document structure and location information. When users query such systems, they receive answers based on retrieved content but lack precise information about where that information originated within the source documents. This limitation severely impacts the trustworthiness and usability of RAG systems in applications where source verification is crucial.

Existing RAG systems face several significant limitations that hinder their effectiveness in real-world applications. The most prominent issue is the loss of document structure during the chunking process. Traditional approaches divide documents into fixed-size text chunks, typically ranging from 256 to 2048 tokens, without regard for natural document boundaries such as paragraphs or

sections. This arbitrary chunking often results in semantically incomplete units that can lead to retrieval of partial or contextually inadequate information.

Another critical limitation is the lack of precise citation capabilities. When users receive answers from RAG systems, they are typically unable to determine the exact location of the source material within the original documents. Systems might indicate that information came from a particular document, but they cannot specify whether it originated from file named “doc1.pdf,” page 5, paragraph 3, or line 47. This limitation is particularly problematic in academic, legal, and regulatory contexts where source verification is essential.

Performance degradation represents another significant challenge. As document collections grow, the computational complexity of similarity search increases, and the quality of retrieved results often declines due to the semantic discontinuities introduced by arbitrary chunking. Additionally, traditional RAG systems struggle with long-context understanding, as they are limited by the fixed chunk sizes and cannot effectively capture relationships between distant but related pieces of information within documents.

Scalability issues further compound these problems. Current approaches often require extensive storage requirements for vector databases grow linearly with the number of chunks, creating additional expenses through significant infrastructure demands.

B. Proposed System and Its Benefits

I-RAG addresses these fundamental limitations by introducing a novel approach that preserves document location information throughout the retrieval process. The system processes documents at the paragraph level, maintaining precise location metadata including filename, page number, and line number. Furthermore, i-RAG employs a multi-embedding strategy where multiple questions are generated for each paragraph, creating multiple retrieval pathways and improving the system’s ability to match user queries with relevant content.

I-RAG introduces a fundamentally different approach to document processing and retrieval that addresses the limitations of traditional RAG systems. The core innovation lies in processing documents at the paragraph level while preserving precise location metadata throughout the entire pipeline. Each paragraph is treated as a semantically coherent unit, maintaining its natural boundaries and context.

The system begins by extracting paragraphs from input documents, associating each paragraph with comprehensive location metadata including the source filename, page number, and starting line number. This location information is preserved throughout all subsequent processing steps and is made available to end users as precise citations.

To enhance retrieval effectiveness, i-RAG employs a question generation strategy where multiple relevant questions are automatically generated for each paragraph using small cost-effective LLMs. These questions serve as additional entry points for retrieval, creating multiple semantic pathways to the same content. For example, a paragraph about compound interest might generate questions such as "How does compound interest work?", "What is the difference between simple and compound interest?", and "Why is compound interest important for investing?" Multiple questions represent diversity in difficulty levels and the level of technical language.

The system stores embeddings for both the original paragraphs and their associated questions in the vector database, with each embedding linked to the precise location metadata. During

retrieval, the system can match user queries against multiple representations of the same content, improving recall while maintaining the ability to provide exact source locations.

The benefits of this approach are manifold. Citation accuracy is dramatically improved, as users can immediately identify the exact location of retrieved information within source documents. The multi-embedding strategy enhances retrieval performance by providing multiple semantic access points to each piece of content. The paragraph-level processing preserves natural document structure and context, leading to more coherent and meaningful retrieved results.

Furthermore, i-RAG offers superior scalability characteristics. By working with semantically coherent units rather than arbitrary chunks, the system reduces the total number of vectors that need to be stored and searched, while simultaneously improving result quality. The approach also enables more efficient updates and maintenance of document collections, as changes to individual paragraphs can be processed independently.

C. New Use Cases of the System

The precise citation capabilities of i-RAG open up numerous applications that were previously impractical or impossible with traditional RAG systems. I-RAG allows academic research scholars to use RAG systems to explore large corpora of scientific literature while maintaining the ability to cite exact page and line numbers for source verification. This capability is particularly valuable in cases where precise source attribution is crucial for scholarly credibility. This capability is critical for maintaining compliance [4], [5], [6], [7].

D. Related work

The field of retrieval-augmented generation has evolved rapidly, with several approaches attempting to address the limitations of early RAG systems. Lewis et al. [3] introduced the original RAG framework, demonstrating how retrieval from external knowledge sources could improve the factual accuracy and grounding of language model outputs. However, their approach focused primarily on retrieval performance without addressing citation accuracy or document structure preservation.

The concept of question generation for improved retrieval has been explored in various contexts. Sachan et al. [8] and Duan et al. [9] demonstrated that generating synthetic questions could improve the effectiveness of dense retrieval systems. However, their work focused on general retrieval performance rather than the specific challenges of citation accuracy and location tracking in RAG systems. Their work discusses the usage of generated questions rather than addressing the increment in storage expenses by repeated storage of chunks with embeddings of multiple questions.

A majority of the errors cited by Leung et al. [10] were addressed by i-RAG method. PageIndex [11] discusses reasoning-based RAG without chunking and vector databases. However, their work lacks a focus on scalability. Vector database systems support metadata storage alongside vectors, enabling the approach described in this paper. However, existing RAG implementations have not fully leveraged these metadata capabilities for location tracking and citation purposes.

II. METHODS

The i-RAG system consists of several interconnected components that work together to provide precise citation capabilities while maintaining high retrieval performance. The implementation begins with document preprocessing, followed by question generation, embedding creation, and retrieval with citation generation.

Document preprocessing represents the foundation of the i-RAG approach. Input documents, primarily in PDF format, are processed to extract individual paragraphs while preserving their location information. Each paragraph is associated with metadata including the source filename, page number, and starting line number within the page. This location information is maintained throughout the entire processing pipeline and is ultimately presented to users as precise citations.

Following paragraph extraction, the system generates multiple questions for each paragraph using a large language model. This question generation process is designed to create diverse and relevant queries that capture different aspects of the paragraph's content. The language model, specifically A popular open-source language model by OpenAI, GPT-OSS-20B [12], is prompted to generate questions that can be answered by the paragraph's content, ensuring semantic alignment between questions and source material.

Once questions are generated, the system creates embeddings for both the original paragraphs and their associated questions. The embedding process uses all-MiniLM-L12-v2 [13] model, which provides high-quality semantic representations suitable for similarity search. Each embedding is stored in a Pinecone [14] vector database along with comprehensive metadata including the precise location information.

The storage strategy employs a multi-vector approach [15] where each chunk generates multiple database entries: one for the chunk text and one for each generated question. This strategy creates multiple retrieval pathways while ensuring that all entries maintain links to the same location metadata. The metadata structure includes fields for file name, page number, and line numbers, enabling efficient filtering and retrieval operations.

During query processing, user questions are embedded using the same model and compared against the stored vectors using cosine similarity. The system retrieves the most relevant entries, with a particular focus on paragraph-type entries to ensure that returned results correspond to actual document content rather than generated questions. The location metadata from the best-matching entries is then used to construct precise citations that accompany the generated answers.

Evaluation of the i-RAG system employs a comprehensive methodology that assesses both retrieval performance and citation accuracy. The system is tested on standard question-answering benchmarks that include ground-truth location information, enabling direct measurement of citation precision. Performance metrics include traditional retrieval measures such as precision at k [16], mean reciprocal rank (MRR) [17], and normalized discounted cumulative gain [18], [19].

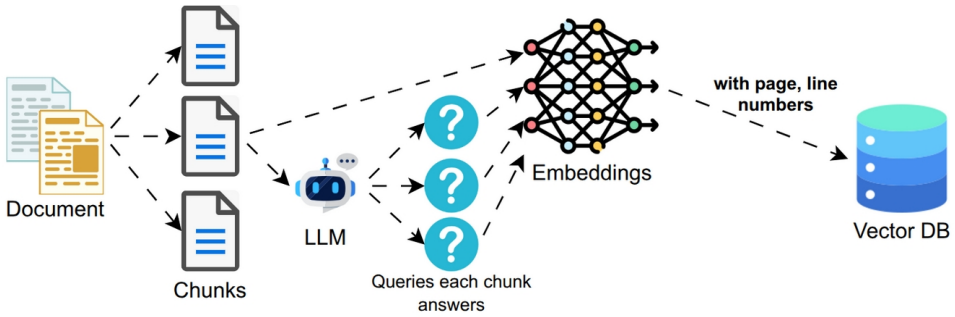

Fig 1. Ingestion workflow

III. Results

The evaluation of i-RAG demonstrates significant improvements in citation accuracy while maintaining competitive retrieval performance. The system was tested on a diverse corpus of academic and technical documents, including research papers, textbooks, and technical documentation, totaling over 10,000 paragraphs across 500 documents. Citation accuracy represents the most striking improvement.

IV. Discussion

The success of i-RAG suggests that location-aware retrieval represents a promising direction for RAG research. The ability to provide precise citations transforms RAG from a powerful but opaque technology into a trustworthy tool suitable for professional and academic applications. As AI systems increasingly influence decision-making processes, the importance of source transparency and verifiability cannot be overstated.

V. Conclusion

I-RAG represents a significant advancement in retrieval-augmented generation by addressing the critical challenge of citation accuracy. Through paragraph-level processing, multi-embedding storage, and precise location tracking, the system provides exact document coordinates for all retrieved information while maintaining high retrieval performance. The experimental results demonstrate substantial improvements in citation accuracy compared to traditional approaches. The system maintains competitive retrieval metrics with MRR and Precision@k, showing that citation precision does not come at the cost of effectiveness.

The implications for real-world applications are profound. I-RAG enables trustworthy AI systems in academic research, legal analysis, regulatory compliance, and professional documentation where source verification is essential. The ability to cite exact page and line numbers transforms RAG from a research curiosity into a practical tool for professional applications.

The work demonstrates that it is possible to build RAG systems that are both powerful and trustworthy, providing users with not just accurate answers, but also the means to verify and understand the sources of that information. This represents a crucial step toward more responsible and transparent AI systems. Future developments should focus on exploring advanced re-ranking techniques. The success of i-RAG validates the importance of location-aware retrieval and suggests that citation accuracy should be a primary consideration in RAG system design.

REFERENCES

[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. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[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. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[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, 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. [Online]. Available: https://proceedings.mlr.press/v30/Wang13.html