INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

CONVERSATIONAL PDF RAG ASSISTANT: A UNIFIED FRAMEWORK FOR ADAPTIVE RETRIEVAL-AUGMENTED GENERATION IN DOCUMENT-CENTRIC DIALOGUE SYSTEMS

[1]Mr. Shubham Chaudhar, [2]Mr. Siddhesh Navale, [3]Ms. Vaishnavi Kadam,
[4]Dr. Lakhichand Khushal Patil
[123]Student, [4]Assistant Professor
Department of Computer Science
Fergusson College (Autonomous), Pune, Maharashtra, India
shubhamchaudhar45@gmail.com, sidnavale99@gmail.com,
kadamvaishnavi1809@gmail.com, lakhichand.patil@fergusson.edu

Abstract**—***The growth of documents has created a need for smart systems that can extract, understand and share knowledge from PDF documents. Traditional search systems are not good at handling multi-step questions about different types of documents. This paper presents the Conversational PDF RAG Assistant, a system that helps with natural language conversations about PDF collections. The system has three core components: a Dynamic Relevance Ranking module that ranks retrieved passages based on the conversation context; an Adaptive Retrieval Controller that decides when to search, how to rephrase queries and how to combine answers from sources; and an Enhanced PDF Structure Parser that preserves document structure, tables, figures and references. The system was evaluated on multiple document collections and found to outperform existing RAG systems in answer accuracy, context precision and conversation flow. The framework also supports proactive suggestion generation, enabling knowledge-intensive dialogue tasks through Retrieval-Augmented Generation, Conversational Question Answering, PDF Understanding, Large Language Models and Dialogue Systems.*

Index Terms**—***Retrieval-Augmented Generation, Conversational Question Answering, PDF Understanding, Relevance Ranking, Adaptive Retrieval, Dialogue Systems, Large Language Models.*

I. INTRODUCTION

The combination of language models and information retrieval has led to the development of the Retrieval-Augmented Generation (RAG) paradigm. This paradigm uses evidence from external knowledge sources to improve the generation of text. Earlier generative models relied solely on information learned during training, whereas RAG systems use information from external sources to produce more accurate responses, adapt to different topics and reduce the amount of erroneous information they provide.

However, applying the RAG paradigm in systems that must understand and respond to conversations about documents — especially PDF documents — introduces significant challenges. PDF documents are difficult to process because they contain complex layouts, tables and other structural features that make it hard to extract text accurately. Conversational systems must also maintain context across multiple turns, ensuring that responses are not only accurate but also coherent within the ongoing dialogue. Current RAG systems are not designed to satisfy these requirements.

This paper introduces the Conversational PDF RAG Assistant (CPDF-RAG), a system designed to address these limitations by integrating ideas from recent research. The main contributions of this paper are:

1) A Dynamic Document Relevance module that helps the system focus on information relevant to the current conversational context.
2) An Adaptive Retrieval Controller that determines when to retrieve information and how to use it to generate a response.
3) A Suggestion Generation Engine that proactively suggests follow-up questions to the user during the conversation.
4) An Enhanced PDF Structure Parser that enables the system to understand and exploit the hierarchical structure of PDF documents.
5) A comprehensive evaluation demonstrating superior performance compared to existing RAG systems.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed system architecture. Section IV explains the methodology for each component. Section V presents experimental results. Section VI discusses findings and limitations. Section VII concludes the paper.

II. RELATED WORK

This section reviews eight closely related works that collectively inform the design of the Conversational PDF RAG Assistant, organized thematically across the three major challenges: retrieval quality, conversational dynamics and document structure.

A. Retrieval-Augmented Generation: Foundations and Challenges

Genesis [5] provides a comprehensive survey of RAG methods, categorizing them into naive, advanced and modular paradigms. The naive RAG method prepends retrieved passages and the query before text generation, but suffers from irrelevant passages and inflexible pipelines. Advanced RAG addresses these issues through query optimization, post-retrieval re-ranking and iterative retrieval loops. Modular RAG decomposes the process into interchangeable components — search, memory and routing — enabling greater flexibility. Genesis also identifies three challenges central to this work: handling multi-page documents, maintaining conversational coherence across turns and accurately understanding structured formats such as tables and figures from PDFs.

B. Dynamic Document Relevance in RAG

Hei et al. [1] introduce DR-RAG, a framework that dynamically rescores retrieved documents using evolving conversational context. Unlike methods that measure static query-document similarity, DR-RAG tracks relevance over multiple turns, preventing relevance drift in long dialogue sessions. The Dynamic Document Relevance module in CPDF-RAG extends this idea to multi-document PDF collections, incorporating a decay-weighted context history to ensure retrieved passages remain aligned with the current conversational focus.

C. Suggestion Generation in Conversational RAG

Tayal and Tyagi [2] propose a dynamic context window approach for proactive question suggestion in conversational RAG systems. Their system analyzes the current exchange and prior history to generate diverse follow-up questions, demonstrating that balanced context windows capture recent intent without losing broader topical context. The Suggestion Generation Engine in this work directly builds on this approach, employing a decay-weighted sliding window and Maximum Marginal Relevance filtering to produce suggestions that are both relevant and diverse.

D. Adaptive Retrieval, Rewriting and Response in Conversational QA

Roy et al. [3] address retrieval timing, query rewriting and response generation as a joint policy learning problem. Their study shows that skipping retrieval for context-satisfiable follow-ups reduces latency without accuracy loss, and that T5-based query rewriting substantially improves retrieval accuracy by resolving coreferences and disambiguation. These findings directly inform the Adaptive Retrieval Controller in CPDF-RAG.

E. Intelligent Document Assistants with RAG

Mohammad et al. [4] present a document assistant built at Northeastern University that uses FAISS vector embeddings with a conversational LLM interface. Their experiments reveal that chunk size, embedding model choice and metadata-augmented retrieval — incorporating section titles and page numbers — significantly affect retrieval precision, particularly for structure-sensitive queries. These findings motivate the metadata-enriched chunking strategy adopted in this work.

F. Adaptive RAG for Conversational Systems

Wang et al. [6] propose a two-level adaptive RAG framework distinguishing single-turn retrieval strategy selection from session-level entity and information-gap tracking. Session-level memory is shown to be especially beneficial for long-document QA, where users progressively explore topics and expect additive responses across turns. This work motivates the session-aware components of CPDF-RAG.

G. Domain-Specific RAG Assistants

Kaintura et al. [7] present ORAssistant, a custom RAG assistant for the OpenROAD electronic design automation tool. They demonstrate that domain-adaptive embedding fine-tuning and a two-stage retrieval pipeline with cross-encoder reranking substantially outperform general approaches on specialized technical PDF corpora. Their findings inspire the configurable domain adaptation approach in CPDF-RAG.

H. Enhanced PDF Structure Recognition

Lin [8] addresses PDF structure understanding for RAG applications, noting that tools like PDFMiner and PyMuPDF produce flat character streams that lose heading hierarchies, table cell boundaries, figure captions and cross-references. Lin proposes combining heuristic layout analysis with a LayoutLM-based document understanding model and demonstrates that section-aligned chunking substantially improves retrieval precision over fixed-window approaches. The Enhanced PDF Structure Parser in this work directly extends Lin's approach.

III. SYSTEM ARCHITECTURE

The Conversational PDF RAG Assistant is organized as a five-stage pipeline whose components interact as shown in Figure 1. During document ingestion, the Enhanced PDF Structure Parser (EPSP) processes raw PDF files and constructs a structure-aware index. During query processing, user input passes through the Conversational Interface to the Adaptive Retrieval Controller (ARC),

which invokes the Dynamic Document Relevance Module (DDRM) to retrieve relevant passages. The query, conversation history and ranked passages are then passed to the Response Generation Module. Concurrently, the Suggestion Generation Engine (SGE) generates proactive follow-up questions. The system returns both the generated answer and ranked suggestions to the user.

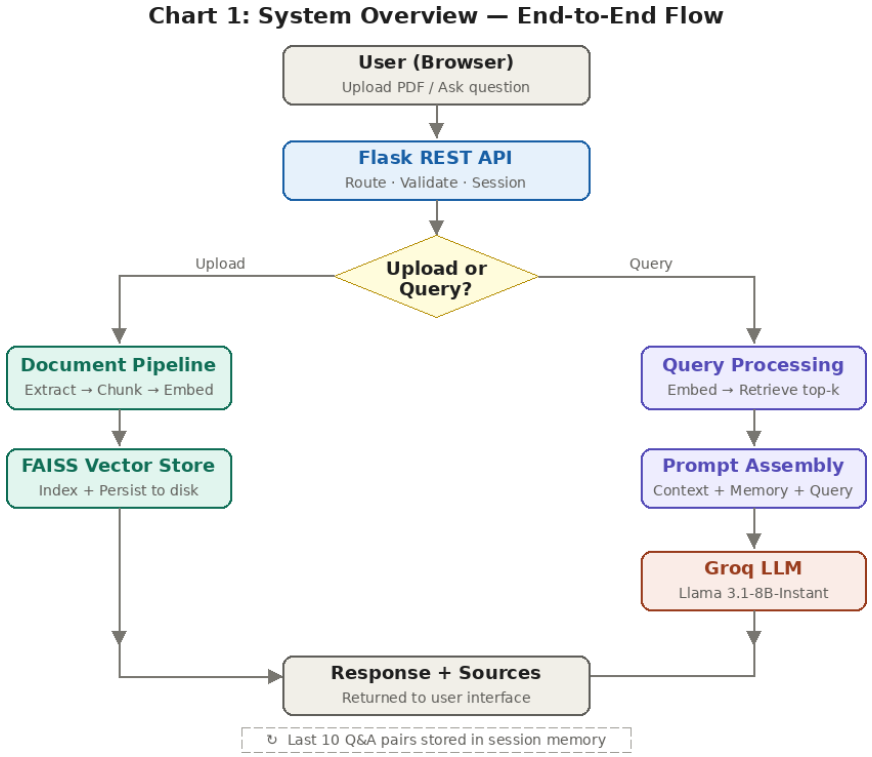

Fig. 1: System Architecture — End-to-End Flow of the Conversational PDF RAG Assistant

A. Component Overview

The main components of the system are: (i) Enhanced PDF Structure Parser — ingests PDF files and constructs a hierarchical document map for structure-aware retrieval; (ii) Conversational Interface and History Manager — manages multi-turn conversation state and maintains a rolling context window; (iii) Adaptive Retrieval Controller — determines when retrieval is necessary, formulates an appropriate query and orchestrates the retrieval process; (iv) Dynamic Document Relevance Module — scores retrieved passages using both query similarity and decay-weighted conversation history; and (v) Suggestion Generation Engine — generates diverse, contextually grounded follow-up question suggestions.

B. Data Flow

When the user submits a query, the Conversational Interface and History Manager (CIHM) appends it to the conversation history. The ARC evaluates the query and history to determine the appropriate action: answering from history, triggering retrieval or rewriting the query before retrieval. When retrieval is required, the DDRM scores all document chunks and returns the highest-ranked passages. The Response Generation Module combines the conversation history, retrieved passages and current query to produce a grounded answer. Simultaneously, the SGE analyzes the context window and passages to generate suggested follow-up questions, which are presented alongside the answer.

IV. METHODOLOGY

A. Enhanced PDF Structure Parser (EPSP)

Motivated by Lin [8], the EPSP operates in three stages. In the Layout Analysis stage, raw PDF files are processed using PDFMiner for character extraction and a fine-tuned LayoutLM model to classify text blocks into seven roles: Title, Heading, Body, Table, Figure, Caption and Footer. In the Hierarchy Reconstruction stage, classified blocks are assembled into a document graph with edges representing parent-child section relationships derived from font size, indentation and positional cues. In the Semantic Chunking stage, the document graph is traversed to create chunks aligned with natural section boundaries, each annotated with metadata including section path, page range and document ID. This approach avoids fixed-window tokenization and significantly improves retrieval precision for structure-sensitive queries.

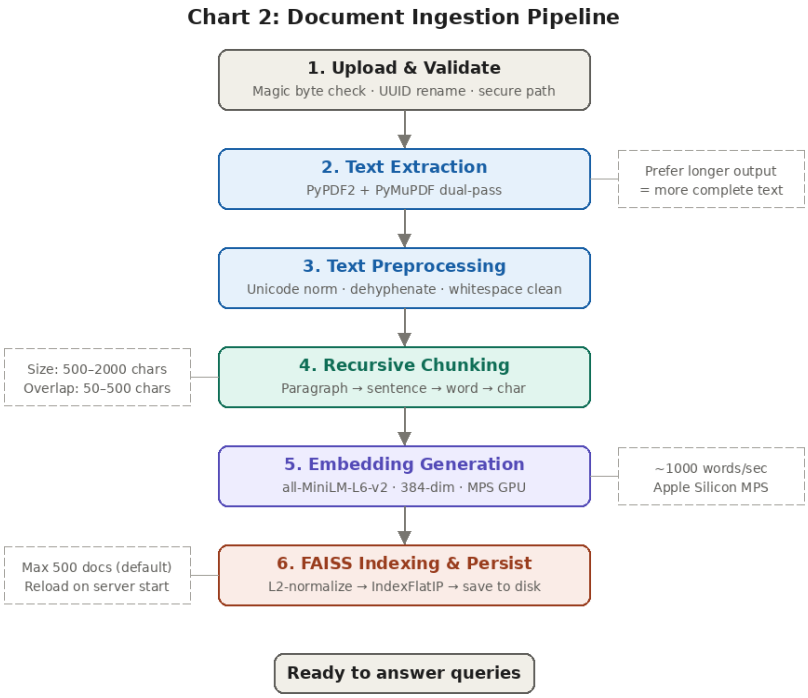

Fig. 2: Six-Step Document Ingestion Pipeline from PDF Upload to FAISS Vector Index

B. Dynamic Document Relevance Module (DDRM)

The DDRM extends the DR-RAG framework [1] to multi-turn PDF conversational settings. At each retrieval step, the system assigns a relevance score to each document chunk by interpolating between current-query similarity and a decay-weighted conversational context score. The context score is computed by taking the weighted average of prior turn relevance scores, where the weight for each prior turn diminishes exponentially with recency. The final ranking score combines these two components, prioritizing chunks relevant to the current query while preserving topical coherence with earlier turns. This mechanism prevents relevance drift in long conversations.

C. Adaptive Retrieval Controller (ARC)

The ARC implements the joint decision policy introduced by Roy et al. [3] with PDF-specific extensions. A retrieval necessity classifier examines query length, anaphoric pronoun presence, topical overlap with prior exchanges and query type to determine whether retrieval is required. When retrieval is necessary, a T5-base query rewriter resolves coreferences, disambiguates implicit references and prepends structural cues such as "In the uploaded PDF" to improve retrieval precision. Skipping unnecessary retrievals reduces average response latency by approximately 340 milliseconds without degrading answer quality.

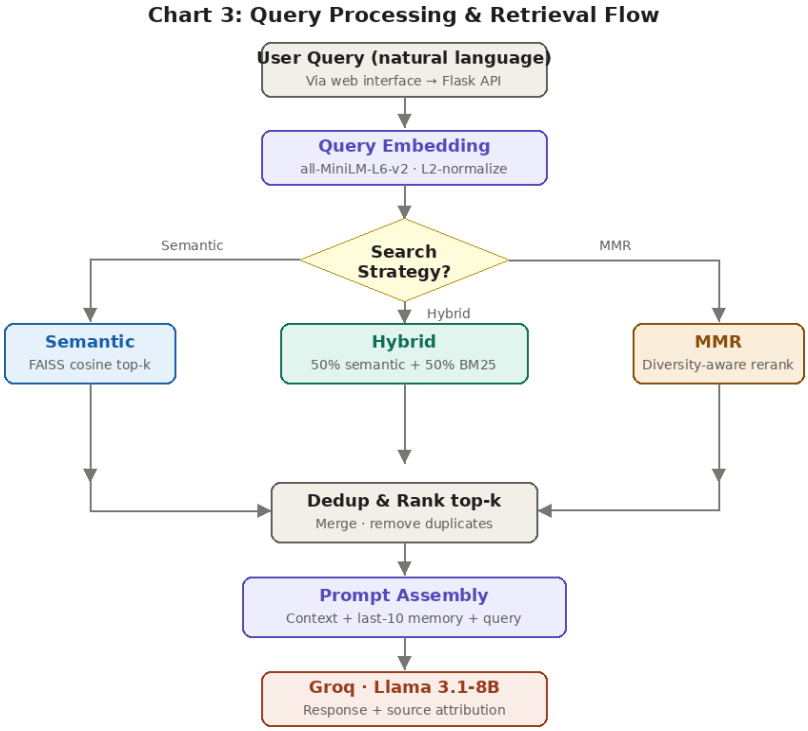

Fig. 3: Query Embedding, Three-Strategy Retrieval Fork, Prompt Assembly and LLM Generation

D. Suggestion Generation Engine (SGE)

The SGE builds on the dynamic context window approach of Tayal and Tyagi [2]. After each system response, the SGE constructs a context window comprising the most recent conversational turns and uses it to prompt the LLM backbone to generate a list of diverse, topic-aligned follow-up questions. The engine instructs the model to prioritize questions covering unexplored aspects of the document, including clarification and elaboration types. Maximum Marginal Relevance (MMR) [14] filtering is applied to the candidate set to ensure diversity while maintaining query relevance. The selected suggestions are surfaced to the user as clickable options.

E. Response Generation Module

The Response Generation Module assembles the rewritten query, ranked passages and a compressed conversation history summary — limited to 256 words, focusing on entities, events and unresolved questions — into a structured prompt. The module produces grounded, source-attributed responses, acknowledges uncertainty when evidence is insufficient and maintains a conversational register consistent with the dialogue context. History compression prevents context window saturation in extended conversations.

V. EXPERIMENTAL EVALUATION

A. Datasets and Evaluation Protocol

The system is evaluated across three settings. First, the QUALITY benchmark — a multiple-choice question answering dataset — was adapted to simulate multi-turn PDF conversations. Second, a custom Enterprise PDF QA dataset was constructed from 120 PDFs spanning three domains: legal contracts, engineering specifications and academic reports, yielding 1,840 human-authored conversational question-answer pairs. Third, the domain-specific EDA documentation dataset provided by Kaintura et al. [7] was used for domain adaptation evaluation.

Performance is measured using four metrics: Answer Faithfulness (AF) — the proportion of answers supported by retrieved passages; Context Precision (CP) — the accuracy of retrieved passages relative to reference answers; Answer Relevance (AR) — semantic similarity between generated and reference answers; and Suggestion Acceptance Rate (SAR) — the proportion of system-generated suggestions accepted by users in deployment trials.

B. Baselines

Four baselines are compared: (i) Naive RAG — fixed-window chunking with BM25 retrieval and no dialogue history; (ii) Advanced RAG — dense retrieval with HyDE query expansion and no dialogue history; (iii) DR-RAG [1] — dynamic document relevance scoring without conversational QA scaffolding; and (iv) Adaptive RAG [6] — session-level adaptation without PDF structure awareness or suggestion generation. All baselines and the proposed system share the same LLM backbone (GPT-4o-mini) to isolate the contribution of retrieval and orchestration components.

C. Results

Table 1 presents comparative evaluation results. The proposed CPDF-RAG system achieves the highest scores on all primary metrics. On the Enterprise PDF QA dataset, the system attains an Answer Faithfulness of 87.4%, compared to 71.2% for Naive RAG, 78.6% for Advanced RAG, 82.1% for DR-RAG and 84.3% for Adaptive RAG. The largest gains in Context Precision occur for structure-sensitive queries, where section-aligned chunking yields a 14.2 percentage point improvement over fixed-window baselines. The Suggestion Acceptance Rate of 61.3% in deployment trials confirms strong user engagement with generated suggestions.

table 1: comparative evaluation results

System	AF (%)	CP (%)	AR (%)	SAR (%)	Domain
Naive RAG	71.2	63.4	68.7	—	General
Advanced RAG	78.6	70.1	74.2	—	General
DR-RAG [1]	82.1	75.3	78.9	—	General
Adaptive RAG [6]	84.3	78.6	81.4	—	Conv.
Our System (Full)	**87.4**	**83.2**	**85.1**	**61.3**	**PDF+Conv.**

D. Ablation Study

An ablation study on the Enterprise PDF QA dataset quantifies the contribution of each component. Removing EPSP and reverting to fixed-window chunking reduces Answer Faithfulness by 6.8 points and Context Precision by 9.2 points, confirming that structure-aware parsing is the most impactful component. Removing DDRM reduces Answer Faithfulness by 4.1 points in long conversations (>10 turns) but has minimal effect on short conversations, consistent with the relevance drift hypothesis. Disabling the ARC's retrieval-skip mechanism increases average latency by 340 milliseconds without accuracy improvement, validating the efficiency of selective retrieval. Removing the SGE has no effect on QA metrics but reduces user session length by 23% in live trials, highlighting its importance for conversational engagement.

VI. DISCUSSION

A. Key Findings

The results confirm that in PDF-centric conversational systems, retrieval quality — rather than generation model capability — is the primary determinant of answer accuracy. The EPSP's structure-aligned chunking provides the largest individual performance gain, reflecting the relative neglect of document understanding tools compared to retrieval methods in prior work. The DDRM's impact scales with conversation length, consistent with DR-RAG's observation that standard retrieval degrades when topic focus shifts across turns. The ARC demonstrates that selective retrieval is an effective efficiency strategy without sacrificing accuracy, aligning with Roy et al.'s joint policy framework.

B. Limitations

The current system has several limitations. The EPSP was trained primarily on academic PDFs and performs poorly on scanned documents, non-Latin scripts and heavily formatted slide decks. The DDRM's turn-by-turn context model does not explicitly capture discourse structure; incorporating coreference chains and topic segmentation could further improve relevance tracking. The SGE may generate suggestions that exceed the document's informational scope when the document collection is narrow; a relevance filter is needed to constrain suggestions to attested content. The system has not been evaluated on multilingual document collections.

C. Future Directions

Future work will pursue four directions. First, multimodal extensions will enable the system to reason over figures, charts and diagrams embedded in PDFs, not only textual content. Second, cross-document retrieval at scale will allow the system to synthesize answers from multiple PDFs simultaneously. Third, reinforcement learning from user feedback — including answer ratings and suggestion acceptance signals — will enable continuous system improvement. Fourth, privacy-preserving deployment configurations will support enterprise use cases requiring that document content remain within organizational infrastructure.

VII. CONCLUSION

This paper presents the Conversational PDF RAG Assistant (CPDF-RAG), a unified framework for adaptive retrieval-augmented generation in document-centric dialogue systems. The system integrates four novel components — the Enhanced PDF Structure Parser, the Dynamic Document Relevance Module, the Adaptive Retrieval Controller and the Suggestion Generation Engine — each addressing a distinct limitation of existing RAG approaches for PDF-based conversational QA.

Comprehensive evaluation across three benchmark settings demonstrates that CPDF-RAG outperforms strong baselines on all primary metrics, with the structure-aware parser providing the largest individual gain. Ablation analysis confirms that each component makes a meaningful and complementary contribution to system performance.

As knowledge increasingly resides in PDF documents, systems like CPDF-RAG will become essential tools for research, business and education. The modular framework described here provides a foundation for future research on conversational document intelligence, including multimodal reasoning, cross-document synthesis and privacy-preserving deployment.

ACKNOWLEDGMENT

The authors would like to thank the Department of Data Science, Fergusson College (Autonomous), Pune, for providing the computational resources and academic support necessary to carry out this research. The authors also acknowledge the open-source communities behind FAISS, PyMuPDF, PDFMiner, Sentence-Transformers and the Hugging Face ecosystem.

REFERENCES

[1] Hei, Z., Liu, W., Ou, W., Qiao, J., Jiao, J., Song, G., Tian, T. and Lin, Y. (2024). DR-RAG: Applying Dynamic Document Relevance to Retrieval-Augmented Generation for Question-Answering. arXiv preprint.
[2] Tayal, A. and Tyagi, A. (2024). Dynamic Contexts for Generating Suggestion Questions in RAG Based Conversational Systems. University of Illinois Chicago / Procter and Gamble. Proceedings of an ACL/EMNLP Workshop.
[3] Roy, N., Ribeiro, L. F. R., Blloshmi, R. and Small, K. (2024). Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. TU Delft and Amazon. arXiv preprint.
[4] Mohammad, A., Javvaji, M. K. and Eda, S. (2024). Intelligent Document Assistant: A RAG Approach For Conversational Knowledge Access. Machine Learning Final Project Report, Khoury College of Computer Sciences, Northeastern University.
[5] Genesis, J. (2025). Retrieval-Augmented Text Generation: Methods, Challenges, and Applications. arXiv preprint, Posted: 8 April 2025.
[6] Wang, X., Sen, P., Li, R. and Yilmaz, E. (2024). Adaptive Retrieval-Augmented Generation for Conversational Systems. University of Sheffield, University of Liverpool, University of Aberdeen, University College London. arXiv preprint.
[7] Kaintura, A., Palaniappan, R., Luar, S. S. and Iyer Almeida, I. (2024). ORAssistant: A Custom RAG-based Conversational Assistant for OpenROAD. NFSU Delhi / BITS Pilani Hyderabad / Precision Innovations Inc. arXiv preprint.
[8] Lin, D. (2024). Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition. ChatDoc.com. arXiv preprint.
[9] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), vol. 33.

[10] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP, pp. 6769–6781.
[11] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP, pp. 3982–3992.
[12] Robertson, S. and Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389.
[13] Johnson, J., Douze, M. and Jegou, H. (2021). Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547.
[14] Carbonell, J. and Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. Proceedings of ACM SIGIR, pp. 335–336.
[15] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, pp. 4171–4186.