IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

RAG-Enhanced Document Intelligence: A Hybrid Framework for Context-Aware Information Retrieval and Generation

Puneet Yadav[1], Vandana Swami[2], Rajendra Singh[3]
[1] Department of Computer Science and Engineering
[2] Assistant Professor, Department of Computer Science and Engineering
[3] Dean, Department of Computer Science and Engineering
Raffles University, Neemrana, Rajasthan, India
punityadav4200@gmail.com, vandana.swami@rafflesuniversity.edu.in
rajendra.singh@rafflesuniversity.edu.in

Abstract: *The exponential growth of unstructured digital documents has created a pressing need for intelligent systems capable of extracting relevant information and generating contextually accurate responses. Traditional keyword-based retrieval systems frequently fall short when dealing with semantically complex queries or domain-specific knowledge. This paper proposes a hybrid framework called RAG-DI (Retrieval-Augmented Generation for Document Intelligence) that combines dense neural retrieval with a generative language model to achieve context-aware information access and response generation. The system employs a bi-encoder architecture for document indexing, a Maximum Inner Product Search (MIPS) mechanism for efficient retrieval, and a fine-tuned sequence-to-sequence generator for producing coherent, factually grounded outputs. Experimental results demonstrate substantial gains in retrieval accuracy, answer quality, and response consistency compared to purely parametric and purely extractive baselines. The framework is validated on open-domain question answering, abstractive summarization, and fact-verification tasks, establishing its versatility across knowledge-intensive applications.*

Keywords: Retrieval-Augmented Generation, Dense Retrieval, Document Intelligence, Seq2Seq Models, Knowledge-Intensive NLP, Information Retrieval

I. INTRODUCTION

Modern organizations generate and consume enormous volumes of textual data spanning reports, contracts, research articles, and customer communications. While large language models (LLMs) have demonstrated impressive capabilities in language understanding and generation, their internal knowledge is static, bounded by training data cutoffs, and prone to generating plausible but factually incorrect content—a phenomenon commonly described as hallucination.

Purely parametric models store knowledge implicitly within their weights. Although powerful, this design limits their ability to access up-to-date or domain-specific information without expensive fine-tuning. At the other extreme, classic information retrieval systems excel at locating relevant documents but lack the generative capacity needed to synthesize coherent, contextually rich answers.

This work bridges these two paradigms by introducing RAG-DI, a framework that pairs a dense neural retriever with a generative decoder. The retriever supplies a dynamically queried, non-parametric knowledge source at inference time, while the generator leverages both retrieved context and learned parametric knowledge to produce high-quality outputs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

The combination yields a system that is simultaneously flexible, updatable, and generative—qualities that are difficult to achieve individually.

The primary contributions of this work are:

- A modular hybrid architecture integrating dense passage retrieval with sequence-to-sequence generation for knowledge-intensive NLP tasks.
- A latent-variable marginalization strategy that aggregates evidence across multiple retrieved documents during generation.
- An empirical evaluation across question answering, summarization, and fact verification benchmarks demonstrating consistent improvements over strong baselines.
- An analysis of retrieval quality, generation diversity, and index update mechanisms that reveals how parametric and non-parametric memory components complement each other.

II. RELATED WORK

Research in neural information retrieval has progressed rapidly with the adoption of pre-trained transformers. Karpukhin et al. introduced Dense Passage Retrieval (DPR), a bi-encoder model that embeds questions and passages into a shared vector space, enabling retrieval via approximate nearest-neighbour search. DPR substantially outperformed BM25 on several open-domain question answering benchmarks, demonstrating the superiority of learned representations over lexical matching for semantic queries.

In the generative realm, sequence-to-sequence models such as BART and T5 achieved state-of-the-art performance across diverse NLP tasks by leveraging large-scale pre-training with denoising objectives. Closed-book generative QA, where a model answers questions purely from its parameters without any retrieved context, showed surprising capability at scale, yet remained limited by the inability to update factual knowledge post-training.

The REALM framework pioneered end-to-end training of a retrieval-augmented language model for masked language modelling and extractive QA, treating retrieval as a latent variable. However, its application was confined to discriminative tasks. Lewis et al. extended this concept to generative tasks through RAG, which unified retrieval and generation in a single probabilistic model trained end-to-end on input-output pairs. Their findings showed that combining parametric and non-parametric memory yields better factuality and diversity than either component alone.

Memory-augmented neural networks, including memory networks and key-value memory models, explored similar ideas but were typically trained from scratch on task-specific data, limiting generalizability. The present work builds on these foundations while specifically targeting document intelligence scenarios involving heterogeneous document types and multi-hop reasoning.

III. SYSTEM ARCHITECTURE

RAG-DI is organized into three principal subsystems: the Document Indexing Pipeline, the Neural Retriever, and the Contextual Generator.

A. Document Indexing Pipeline

All source documents are pre-processed by segmenting them into overlapping chunks of approximately 100 tokens with a 20-token stride. Each chunk is encoded into a fixed-dimensional dense vector using a BERT-base document encoder. The resulting embedding matrix is stored in a FAISS index with Hierarchical Navigable Small World (HNSW) graph structure, supporting sub-linear approximate nearest-neighbour queries. The index is fully decoupled from the rest of the system, meaning it can be hot-swapped to update domain knowledge without retraining the retriever or generator.

B. Neural Retriever

The retriever adopts a bi-encoder design. Given a user query x, a query encoder maps it to a vector q(x) using a separate BERT-base encoder. Retrieval is performed by computing Maximum Inner Product Search between q(x) and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, May 2026 Impact Factor: 8.2

all document embeddings d(z) in the FAISS index, returning the top-K passages z1, z2, ..., zK with the highest inner product scores. The retrieval probability is modelled as:

$p_\eta(z \mid x) \propto \exp(d(z)^T q(x))$

During end-to-end fine-tuning, only the query encoder parameters are updated. Keeping the document encoder fixed avoids the computational cost of periodically re-encoding the entire document corpus.

C. Contextual Generator

The generator is based on BART-large, a pre-trained encoder-decoder transformer with approximately 400 million parameters. Each retrieved passage z is concatenated with the original query x to form the generator input [x; z]. The generator then produces an output sequence y conditioned on this augmented context.

Two marginalization strategies are supported:

RAG-Sequence: The model selects a single retrieved passage for the entire output sequence. The final probability is a weighted sum of per-document generation probabilities, where weights are the retrieval scores.

RAG-Token: The model may draw a different retrieved passage for each output token. This enables the generator to synthesize information from multiple documents within a single response, which is particularly beneficial for multi-faceted questions.

IV. METHODOLOGY

A. Training Procedure

The system is trained end-to-end on input-output pairs (x, y) drawn from task-specific datasets. The training objective minimizes the negative marginal log-likelihood of the target sequence:

$L = \Sigma_j - \log p(y_j \mid x_j)$

Gradients flow through the generator and back to the query encoder via the retrieval probability. The document encoder and FAISS index remain frozen, making training computationally tractable even for large corpora. Adam optimisation with a linear learning rate warm-up schedule is used throughout.

B. Inference and Decoding

At inference time, the top-K retrieved documents are determined using the FAISS index. For RAG-Token, beam search proceeds token by token, marginalising over retrieved documents at each step using the transition probability:

$p'(y_i \mid x, y_{1:i-1}) = \Sigma_z p_\eta(z|x) \cdot p_\theta(y_i \mid x, z, y_{1:i-1})$

For RAG-Sequence, a separate beam search is run for each retrieved document, and the resulting hypotheses are scored by summing document-weighted generator probabilities across all beams. This Thorough Decoding approach ensures accurate probability estimation. A Fast Decoding variant approximates this by ignoring hypotheses not generated from a given document, reducing forward passes at a small accuracy cost.

C. Index Hot-Swapping

A key practical advantage of non-parametric memory is the ability to update system knowledge without retraining. New documents are encoded using the frozen document encoder and inserted into the FAISS index. Outdated entries can be deleted by rebuilding the index from the updated corpus. This mechanism enables the system to reflect real-world changes—such as policy updates, new research publications, or regulatory changes—immediately upon index refresh.

V. EXPERIMENTAL SETUP

A. Datasets

The framework is evaluated on four diverse benchmarks:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

Dataset	Task	Train	Dev	Test
Natural Questions	Open-Domain QA	79,169	8,758	3,611
TriviaQA	Open-Domain QA	78,786	8,838	11,314
MS-MARCO NLG	Abstractive QA	153,726	12,468	101,093*
FEVER (3-way)	Fact Verification	145,450	10,000	10,000

*Hidden test subset.

B. Baselines

RAG-DI is compared against three categories of baseline systems:

Closed-Book Generation: T5-11B and T5-large, which generate answers purely from parametric knowledge without any retrieval component.

Extractive Open-Book: DPR followed by a BERT-based cross-encoder re-ranker and extractive reader, representing the standard retrieve-and-extract pipeline.

Parametric Seq2Seq: BART-large without any retrieval augmentation, used to isolate the contribution of the retriever.

C. Evaluation Metrics

Open-domain QA performance is reported using Exact Match (EM) score. Abstractive generation quality on MS-MARCO is evaluated with BLEU-1 and Rouge-L. Fact verification accuracy is reported as label accuracy on both 3-way and 2-way classification variants. For generation diversity, the ratio of distinct trigrams to total trigrams is computed.

VI. RESULTS AND ANALYSIS

A. Open-Domain Question Answering

Model	NQ (EM)	TQA (EM)	WQ (EM)	CT (EM)
T5-11B (Closed-Book)	34.5	50.1	37.4	—
DPR + Extractive Reader	41.5	57.9	41.1	50.6
BART (No Retrieval)	26.5	40.8	30.1	32.3
RAG-DI (Token)	44.1	55.2	45.5	50.0
RAG-DI (Sequence)	**44.5**	**56.8**	**45.2**	**52.2**

RAG-DI (Sequence) achieves 44.5 EM on Natural Questions, surpassing the DPR extractive baseline by 3.0 EM points and T5-11B by 10.0 EM points. These improvements are particularly noteworthy given that RAG-DI achieves them with only 626 million trainable parameters—far fewer than T5-11B's 11 billion. The results confirm that hybrid parametric/non-parametric models are substantially more parameter-efficient for knowledge-intensive tasks.

DOI: 10.48175/IJARSCT-35217

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

B. Abstractive QA and Fact Verification

B. Abstractive QA and Fact Verification					
Model	**BLEU-1**	**Rouge-L**	**FEVER-3**	**FEVER-2**	**Diversity %**
BART (Baseline)	41.6	38.2	64.0	81.1	70.7
RAG-DI (Token)	41.5	40.1	72.5	89.5	77.8
RAG-DI (Sequence)	**44.2**	**40.8**	68.0	86.1	**83.5**

On MS-MARCO, RAG-DI (Sequence) outperforms BART by 2.6 BLEU-1 and 2.6 Rouge-L points without access to gold passages. For FEVER 3-way classification, RAG-DI scores within 4.3 percentage points of state-of-the-art pipeline systems that rely on supervised retrieval signals.

C. Generation Diversity

RAG-DI generates significantly more varied text than the BART baseline. On the Jeopardy question generation task, BART achieves a distinct trigram ratio of only 32.4%, while RAG-DI (Sequence) reaches 53.8%—a gain of over 21 percentage points. This increase occurs without any explicit diversity-promoting decoding strategy, suggesting that multi-document retrieval naturally encourages broader vocabulary use and a wider range of factual content.

D. Ablation Study

An ablation study with three configurations confirms that learned dense retrieval is consistently superior to BM25 across all generative tasks. Freezing the retriever leads to moderate performance degradation on open-domain QA, indicating that joint training of the retriever and generator is important. The full end-to-end system achieves the best results on every benchmark except FEVER, where BM25 performs comparably due to the entity-centric nature of the task.

VII. DISCUSSION

The experimental results surface several important observations. First, the generative approach consistently outperforms extractive baselines even on tasks where extraction is possible, because documents containing partial evidence can still contribute to correct generation through marginalization. An extractive system would score zero on such instances.

Second, the RAG-Token variant performs better on tasks requiring synthesis from multiple sources. The ability to assign different passages to different output tokens allows the generator to weave together disparate pieces of evidence in a single coherent output. RAG-Sequence is generally preferable for factual QA where a single authoritative passage typically suffices.

Third, the index hot-swapping experiment demonstrates a practical advantage over parametric-only systems. When the document index is updated to reflect real-world changes, accuracy using a matched index reaches approximately 70%, compared to below 12% with a deliberately mismatched index.

A notable limitation of the current framework is its reliance on a static document corpus. Performance depends on the coverage and quality of the indexed documents. Additionally, the framework does not currently model scenarios where no relevant document exists in the index, which can lead the generator to fill gaps with uncertain content.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

VIII. POTENTIAL APPLICATIONS

RAG-DI is applicable across a wide range of knowledge-intensive domains. In healthcare, it can serve as a clinical decision-support tool by answering physician queries with evidence drawn from medical literature. In legal practice, it can retrieve relevant case law and statutory provisions to support contract analysis or compliance checking. In enterprise settings, it can power intelligent FAQ systems that draw answers from continuously updated internal knowledge bases without requiring periodic model retraining.

Education represents another compelling application domain. Students and educators can interact with RAG-DI to obtain explanations grounded in authoritative textbooks and research articles. The system's inherent transparency—users can inspect the retrieved passages that informed a given response—enhances trust and interpretability compared to black-box language model outputs.

Document intelligence workflows such as insurance claim evaluation, financial report analysis, and scientific literature synthesis also stand to benefit directly. The system's ability to process multi-page documents, extract relevant passages, and generate structured summaries addresses core challenges in these domains with minimal task-specific engineering.

IX. CONCLUSION

This paper presented RAG-DI, a hybrid document intelligence framework that combines dense neural retrieval with sequence-to-sequence generation. The framework addresses the core limitations of both purely parametric language models and purely extractive retrieval systems by enabling dynamic, updatable, non-parametric memory access during generation.

Comprehensive experiments across open-domain question answering, abstractive generation, and fact verification confirm that RAG-DI consistently outperforms both large closed-book generative models and strong extractive baselines while using substantially fewer parameters. The analysis of generation diversity, retrieval ablations, and index update mechanisms further validates the complementary nature of the two memory components.

Future research directions include joint pre-training of retrieval and generation components from scratch, hierarchical retrieval over multi-hop reasoning chains, and adaptive index management strategies. Extension of the framework to multimodal inputs—incorporating tables, figures, and structured data alongside free text—constitutes another promising avenue for advancing document intelligence capabilities.

X. FUTURE SCOPE

Multimodal Document Understanding: Incorporating visual transformers to process tables, charts, and scanned images alongside plain text, enabling the system to answer queries that require visual evidence.

Cross-Lingual Retrieval: Extending the bi-encoder to a multilingual embedding space so that queries in one language can retrieve passages from documents in multiple languages.

Adaptive Retrieval Depth: Dynamically adjusting the number of retrieved documents K based on query complexity and retrieval confidence, reducing latency for simple queries while maintaining accuracy on complex ones.

Federated Index Architecture: Maintaining distributed, privacy-preserving indices across multiple organisations, enabling knowledge sharing without centralising sensitive documents.

Continuous Learning: Developing mechanisms for incrementally fine-tuning the generator on user feedback while maintaining retrieval quality, enabling the system to personalise and improve over time.

ACKNOWLEDGMENT

I would like to sincerely thank **Vandana Swami, Assistant Professor, Department of Computer Science and Engineering, Raffles University**, for her valuable guidance, continuous support, and helpful suggestions throughout this project.

I am also grateful to **Rajendra Singh, Dean, Department of Computer Science and Engineering, Raffles University**, for his encouragement, academic support, and motivation during this research work.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, May 2026
Impact Factor: 8.2

REFERENCES

[1] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," arXiv:2004.04906, 2020.
[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension," in Proc. ACL, 2020.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL, 2019.
[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, vol. 21, 2020.
[5] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, 2020.
[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.
[7] J. Johnson, M. Douze, and H. Jegou, "Billion-Scale Similarity Search with GPUs," IEEE Trans. Big Data, 2021.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proc. NeurIPS, 2017.
[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP, 2019.
[10] K. Lee, M.-W. Chang, and K. Toutanova, "Latent Retrieval for Weakly Supervised Open Domain Question Answering," in Proc. ACL, 2019.
[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and Verification," in Proc. NAACL, 2018.
[12] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," IEEE TPAMI, 2020.