WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI Document Retrieval & Comprehension System

[1]Sailesh R, [2]Subiksha S, [3]Yamini R, [4]Mrs. Shakthipriya

[1]Artificial Intelligence and Data Science (Third Year) Sri Shakthi Institute of Technology and Engineering, Coimbatore
[2]Artificial Intelligence and Data Science (Third year) Sri Shakthi Institute of Engineering and Technology,Coimbatore
[3]Intelligence and Data Science (Third year) Sri Shakthi Institute of Engineeringand Technology, Coimbatore
[4]P(Assistant professor) Department of Artificial Intelligence and Data Science Sri Shakthi Institute of Engineering and Technology, Coimbatore
[5]Naveenkumar K(Author) Artificial 5Intelligence and Data Science (Third year) Sri Shakthi Institute of Engineering and Technology, Coimbatore

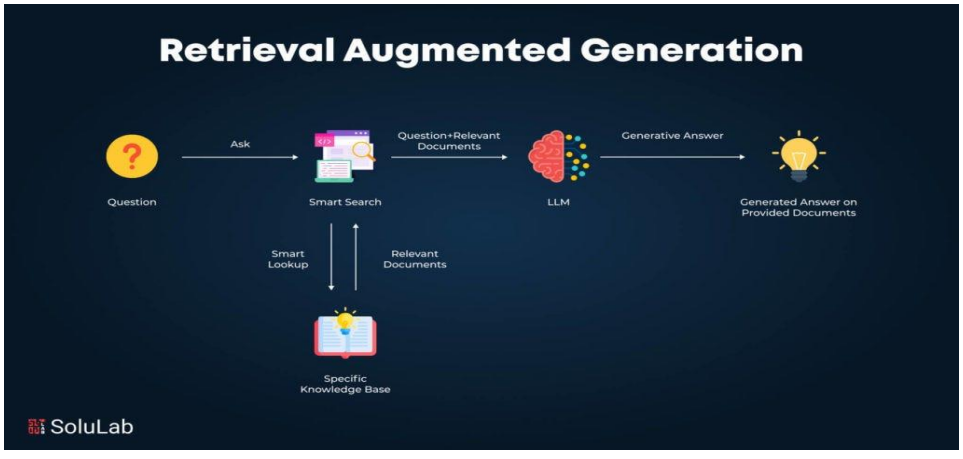
ABSTRACT: -

The AI Document Retrieval and Comprehension System is designed to efficiently locate, extract, and interpret relevant information from large collections of documents. By leveraging natural language processing (NLP), machine learning algorithms, and semantic search techniques, the system enhances the accuracy and speed of information retrieval beyond traditional keyword-based methods. It not only identifies the most pertinent documents based on user queries but also comprehends the context, summarizes key insights, and presents structured answers. This approach reduces the time and effort required for manual document review, supports informed decision making, and can be applied across various domains such as legal research, academic studies, healthcare, and enterprise knowledge management. The system represents a significant step toward making complex information easily accessible and understandable through AI-driven automation.

INTRODUCTION

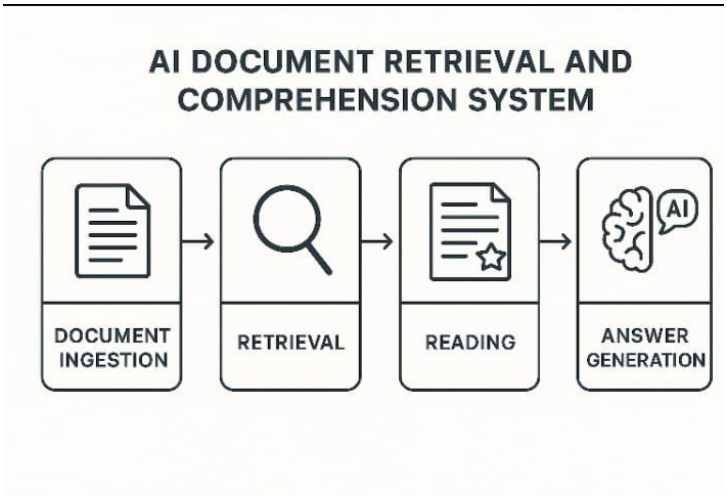
An AI-based document retrieval and comprehension system represents a significant advancement in information management and knowledge extraction. Traditional keyword-based search methods often fall short in understanding the context and semantics of user queries, leading to irrelevant or incomplete results. In contrast, AI-powered systems utilize natural language understanding (NLU), deep learning models like transformers (e.g., BERT, GPT), and semantic embeddings to grasp the intent behind queries and locate the most pertinent information, even if it is expressed in varied language. These systems not only retrieve documents with high precision but also interpret and summarize content, extract specific answers, and highlight key insights within complex texts. This makes them invaluable in environments where rapid, accurate information retrieval is critical—such as in legal case analysis, medical diagnostics, academic research, and enterprise knowledge management. Moreover, they can continuously learn and adapt to new data and evolving user needs, offering a scalable and intelligent approach to handling ever-growing information repositories.

Expanding further, AI-based document retrieval and comprehension systems also integrate various components of artificial intelligence such as information retrieval algorithms, machine reading comprehension, named entity recognition (NER), and context-aware dialogue systems to provide more interactive and human-like responses. These systems are increasingly being deployed in customer support chatbots, digital assistants, and enterprise search engines, where they help users quickly find answers without needing to manually sift through extensive documentation. Additionally, they are capable of handling multilingual data, enabling global accessibility and cross-lingual search capabilities.
One of the key benefits of these systems is their ability to operate over both structured databases and unstructured data sources like PDFs, web pages, and scanned documents using OCR (Optical Character Recognition). This versatility allows organizations to unlock hidden value in legacy documents and massive archives. With continual improvements in pre-trained language models and fine-tuning methods, the accuracy and contextual relevance of AI responses have reached near-human levels in many specialized domains. Furthermore, the integration of feedback loops and user interaction helps these systems refine their understanding over time, leading to better performance and user satisfaction. As data volumes continue to grow exponentially, such intelligent systems are not just enhancements—they are becoming essential tools for efficient knowledge discovery, compliance management, and strategic decision-making.



LITERATURE SURVEY

Artificial Intelligence (AI)-driven document retrieval and comprehension systems are designed to automatically locate and understand relevant information within large corpora of unstructured or semi-structured text. These systems combine natural language processing (NLP), information retrieval (IR), and machine reading comprehension (MRC) techniques. They are used in domains such as legal document analysis, healthcare records processing, customer support, and enterprise knowledge management.



1. Evolution of Document Retrieval: Early document retrieval systems relied on keyword-based search models like TF-IDF and BM25 (Robertson & Zaragoza, 2009). These methods are efficient but lack semantic understanding.
The introduction of word embeddings (e.g., Word2Vec by Mikolov et al., 2013; GloVe by Pennington et al., 2014) allowed for semantic similarity-based retrieval. However, they still lacked contextual awareness.
Breakthroughs in contextual language models, especially BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018), revolutionized retrieval. BERT-based models (e.g., Dense Passage Retrieval, Karpukhin et al., 2020) enabled the matching of questions to relevant document passages with deeper contextual understanding.

2. Machine Reading Comprehension (MRC): MRC tasks require systems not just to retrieve documents, but to understand and extract answers. Early models like BiDAF (Seo et al., 2016) laid the foundation. Transformer-based models (e.g., RoBERTa, ALBERT, ELECTRA) showed strong performance in benchmarks like SQuAD, Natural Questions, and HotpotQA.
Open-domain QA systems (e.g., DrQA by Chen et al., 2017) combine IR and MRC to answer questions using unstructured corpora like Wikipedia. Later systems like RAG (Lewis et al., 2020) and FiD (Izacard & Grave, 2020) improved generation-based comprehension using retriever-generator architectures.

5. Neural IR and Retrieval-Augmented Generation (RAG):
Recent models focus on dense vector representations for semantic retrieval:

- DPR (Dense Passage Retrieval): learns embeddings to retrieve relevant passages.
- ColBERT (Khattab & Zaharia, 2020): allows late interaction for efficiency and effectiveness.
- Contriever: uses unsupervised contrastive learning for retrieval without labeled data.

RAG-based systems (e.g., ChatGPT with retrieval plugins) integrate a retrieval step into generative models like GPT or T5, enabling retrieval-augmented generation, useful in question answering and summarization.

6. Applications and Use Cases:

- Legal AI: LexisNexis and CaseText use NLP for case law retrieval.
- Healthcare: Systems like IBM Watson retrieve and understand medical documents for decision support.
- Enterprise Search: Tools like Haystack (by deepset) and Microsoft Semantic Kernel provide modular pipelines for document comprehension in enterprises.

7. Challenges:

- Scalability: Efficient indexing and retrieval from massive corpora.
- Explainability: Difficulty in understanding why a model retrieved a document.
- Multimodality: Combining text with tables, images, and metadata.
- Bias and Fairness: Ensuring responses are unbiased and contextually fair.
- Continual Learning: Adapting to dynamic and evolving document sets.

8. Future Directions:

- Multilingual and Cross-lingual Retrieval
- Long-document understanding (e.g., Longformer, BigBird)
- Neural symbolic systems that combine logic with deep learning
- Conversational QA systems for dynamic comprehension
- Personalized and contextual retrieval tailored to user profiles

METHODOLOGY

This study presents the design and implementation of an AI-driven Document Retrieval and Comprehension System. The methodology is divided into five major phases: Data Collection, Preprocessing, Document Retrieval, Machine Reading Comprehension, and Evaluation.

1. Data Collection

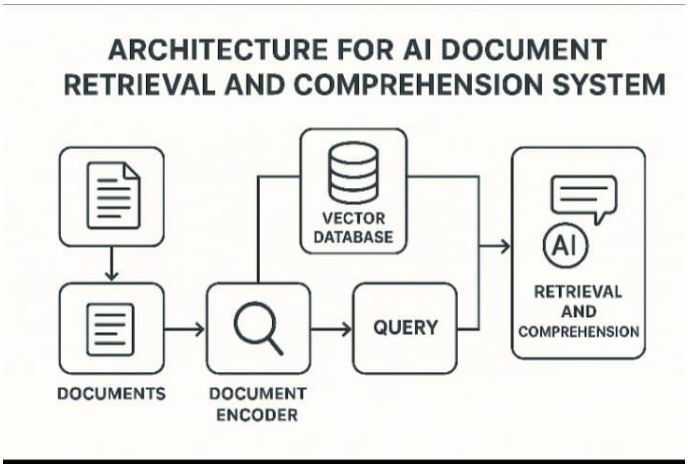
A large-scale unstructured document corpus is used, comprising:

- Open-domain datasets (e.g., Wikipedia, Natural Questions, SQuAD, HotpotQA).
- Domain-specific documents (e.g., legal, healthcare, enterprise records) depending on the use case.

Each document is split into passages or paragraphs to facilitate efficient retrieval.

2. Data Preprocessing

- Text Normalization: Tokenization, lowercasing, stopword removal.
- Segmentation: Long documents are segmented into smaller passages (e.g., 100–300 words).
- Indexing: The passages are indexed using either sparse or dense vector techniques.



3. Document Retrieval System

Two primary approaches are compared:

a. Sparse Retrieval (Baseline)

- Based on BM25 using term-frequency statistics.
- Implemented via tools like Elasticsearch or Lucene.

b. Dense Retrieval (AI-based)

- Embedding-based retrieval using models such as:
    - DPR (Dense Passage Retriever)
    - ColBERT
    - Contriever

Each document and query is encoded into dense vectors using pre-trained transformers (e.g., BERT, RoBERTa). Cosine similarity or dot product is used to retrieve top-k relevant documents.

4. Machine Reading Comprehension (MRC)

An MRC model is employed to extract or generate answers from retrieved passages:

a. Extractive MRC

- Fine-tuned BERT-based models (e.g., BERT, ALBERT) identify start and end tokens of answers within the text.

b. Generative MRC

- Models like T5, RAG, or Flan-T5 are used to generate answers in natural language form from top-k passages.

5. System Integration
The pipeline follows a Retriever-Reader Architecture:
1. Query Input → 2. Retriever finds top-k documents → 3. Reader extracts or generates answers → 4. Output
Optional: Use a retrieval-augmented generation model (like RAG) that combines both steps into one process.

6. Evaluation Metrics
Retrieval Performance:
- Recall@k
- Precision@k
- Mean Reciprocal Rank (MRR)

Comprehension Performance:
- Exact Match (EM)
- F1 Score
- BLEU/ROUGE (for generative outputs)

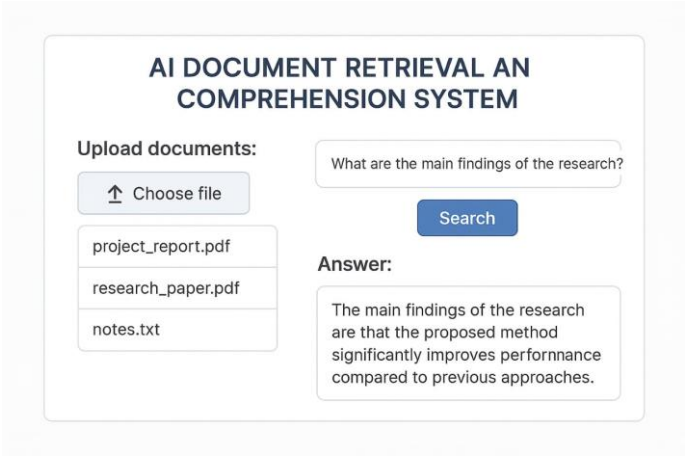
7. Tools and Libraries
- Hugging Face Transformers
- FAISS (for dense vector search)
- Haystack (modular IR/MRC pipelines)
- PyTorch / TensorFlow
- Elasticsearch (for sparse retrieval)

8. Experimental Setup
- Fine-tuning is conducted using GPU-based environments.
- Models are trained and validated using cross-validation on QA datasets.
- Baseline (BM25) vs. neural methods are compared.

CONCLUSION

AI-powered document retrieval and comprehension systems represent a significant advancement in the field of information access and natural language understanding. By combining traditional information retrieval techniques with deep learning-based models such as BERT, T5, and Dense Passage Retrieval (DPR), these systems are capable of not only locating relevant documents but also extracting or generating meaningful, context-aware answers from vast unstructured corpora.

This technology has wide-ranging applications across domains like legal research, healthcare, customer support, and enterprise knowledge management. The transition from keyword-based retrieval to semantic, neural retrieval and machine reading comprehension marks a paradigm shift in how machines process and understand human language.



Despite the progress, challenges remain in terms of scalability, explainability, domain adaptation, and multilingual comprehension. Continued research into retrieval-augmented generation, long-document understanding, and hybrid symbolic-neural approaches will further enhance the robustness and usability of these systems.

In conclusion, AI document retrieval and comprehension systems are transforming how knowledge is accessed and utilized, paving the way for more intelligent, responsive, and user-centric information systems.

REFERENCS

1.Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT. https://arxiv.org/abs/1810.04805

2.Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of EMNLP. https://arxiv.org/abs/2004.04906

3.Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2005.11401

4.Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/

5.Hugging Face Transformers Library. https://huggingface.co/transformers/

6.Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. In IEEE Transactions on Big Data.

7.ElasticSearch Documentation – Full-Text Search Engine. https://www.elastic.co/guide/en/elasticsearch/reference/index.html