



International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

CHAT WITH YOUR PDF - MULTI DOCUMENT Q&A WITH SOURCE CITATIONS

¹ A Akhila, ² K Ramya, ³ K Neha, ⁴ C Aravind, ⁵ Y Sujitha Reddy

¹Assistant Professor, ²³⁴⁵Students

Department of CSE(AIML)

Siddhartha Institute of Technology & Sciences, Narapally

akhilaasara@siddhartha.org.in, 23TQ1A6638@siddhartha.co.in, 23TQ1A6612@siddhartha.co.in,
23TQ1A6619@siddhartha.co.in, 23TQ1A6606@siddhartha.co.in

ABSTRACT

The Chat with Your PDF — Multi-Document Q&A with Source Citations system is an intelligent document-question-answering platform that allows users to interact with multiple PDF documents using natural language. Users often need to read large documents such as research papers, textbooks, reports, manuals, policies, and business documents to find specific information. Searching through multiple PDFs manually can be time-consuming and inefficient.

The proposed system uses Natural Language Processing, document processing, embeddings, vector databases, Retrieval-Augmented Generation (RAG), and Generative AI to understand user questions and retrieve relevant information from uploaded PDF documents. The system extracts text from documents, divides it into meaningful chunks, generates vector embeddings, and stores them in a searchable vector database.

When a user submits a question, the system converts the question into an embedding and searches across the indexed documents to identify the most relevant content. The retrieved document sections are then provided as context to a Generative AI model, which generates a natural-language answer based on the retrieved information. This approach helps reduce the dependence on manually reading entire documents.

A major feature of the proposed system is source citation. Along with the generated answer, the system displays references to the document, page number, section, or relevant text segment used to formulate the response. This improves transparency and allows users to verify the generated information against the original PDF documents.

Overall, the system provides a convenient and interactive way to explore multiple documents through a single conversational interface. It can be useful for students, researchers, employees, legal teams, analysts, and organizations that frequently work with large collections of PDF files. Future enhancements can include multilingual document support, table and image understanding, OCR for scanned PDFs, multimodal RAG, document comparison, advanced citation verification, and enterprise access control.

I. INTRODUCTION

PDF documents are widely used for storing and sharing information in educational, research, business, legal, technical, and administrative environments. A single PDF may contain hundreds of pages, making it difficult for users to quickly locate a particular fact or explanation. When several PDFs need to be studied together, the process becomes even more time-consuming.

Traditional document-search methods generally depend on keyword searches. Users need to enter appropriate keywords and manually inspect the search results to understand the surrounding context. Keyword-based search may also fail when the question and the document use different words to express the same concept.

Recent developments in Generative AI and Large Language Models have created new possibilities for document-based question answering. Instead of simply searching for keywords, AI systems can understand natural-language questions and generate conversational responses. However, a language model alone may not contain the specific information present in a user's private documents and may generate unsupported information.

Retrieval-Augmented Generation (RAG) provides a solution by retrieving relevant information from external documents before generating an answer. In the proposed system, uploaded PDFs are processed and indexed so that relevant passages can be retrieved whenever a user asks a question. Multiple documents can be searched together, allowing the system to answer questions that require information from different files.

The main objective of Chat with Your PDF is to provide a reliable and user-friendly document interaction platform. The system combines semantic retrieval with Generative AI and source citations so that users can obtain concise answers while still being able to verify the information in the original documents.

II. LITERATURE SURVEY

Traditional document-management and search systems provide methods for storing, indexing, and retrieving PDF documents. Keyword-based search engines identify documents or pages containing matching terms. Although effective for exact queries, these systems may not understand the semantic meaning of a question or the relationship between concepts appearing in different documents.

Natural Language Processing has been applied to document-question-answering systems to improve the understanding of natural-language queries. NLP techniques can identify entities, keywords, intent, and semantic relationships within user questions. These techniques provide a foundation for retrieving relevant information from large document collections.

Embedding-based semantic search has improved document retrieval by representing text as numerical vectors. Documents and user queries can be converted into embeddings and compared based on semantic similarity. Vector databases support efficient storage and retrieval of these embeddings, making them useful for large document collections.

Retrieval-Augmented Generation combines information retrieval with Generative AI. Instead of generating an answer solely from the language model's internal knowledge, the system retrieves relevant passages and provides them as context to the model. This approach can improve grounding and make it possible to answer questions using private or recently updated documents.

Multi-document question-answering systems extend RAG by retrieving information from multiple files simultaneously. Such systems can identify relevant passages across different documents and combine them into a single response. However, retrieving irrelevant or conflicting passages can affect response quality, making chunking, ranking, metadata filtering, and context selection important.

Recent document AI systems are also exploring OCR, multimodal document understanding, table extraction, citation generation, reranking, and agentic retrieval. These technologies can improve the ability to understand complex PDFs containing text, images, tables, and scanned pages. The proposed **Chat with Your PDF** system combines multi-document retrieval, semantic search, Generative AI, and source citations to provide an interactive and verifiable document-question-answering experience.

III. SYSTEM ANALYSIS

The **Chat with Your PDF** system is designed to allow users to ask natural-language questions across multiple uploaded PDF documents. The system first accepts one or more PDF files through a secure document-upload interface. A PDF-processing module extracts text, page information, headings, and other supported document metadata from each file. Scanned documents can optionally pass through an OCR module to convert images into machine-readable text. The extracted content is cleaned and divided into meaningful chunks while preserving document and page references. An embedding model converts the chunks into vector representations and stores them in a vector database. When the user submits a question, the query-processing module analyzes the question and generates a corresponding embedding. A semantic retrieval engine searches across the indexed documents and returns the most relevant chunks. A reranking module can prioritize the strongest evidence before context is passed to the Generative AI model. The AI model generates an answer using the retrieved document content rather than relying only on its internal knowledge. A citation-generation module attaches document names, page numbers, and relevant source references to the response. Finally, the user can review the answer and open the cited document location to verify the information.

Existing System

Existing PDF-reading and document-search systems generally require users to manually open files and search for keywords or phrases. Standard PDF viewers provide useful search, highlighting, bookmarking, and navigation features but do not usually provide conversational question answering across multiple documents. When users have several PDFs, they may need to search each document individually and compare the results manually. Keyword-based search may fail when the question uses terminology different from the document. Large documents can require considerable time to read before the user can understand the relevant context. General-purpose AI

chatbots may provide conversational answers but may not have access to private PDF files unless documents are explicitly uploaded or connected. Systems that generate answers without retrieval may also produce unsupported information. Some document-question-answering systems provide answers without clear source references, making verification difficult. Traditional tools may have limited support for comparing information across several documents. Scanned PDFs can also be difficult to search without OCR processing. These limitations create a need for a multi-document conversational system with semantic retrieval and source citations.

Disadvantages of Existing System

- Manual reading of large PDF documents is time-consuming.
- Keyword search may not understand question meaning.
- Multiple PDFs may need to be searched separately.
- Difficult to compare information across documents.
- Limited conversational interaction.
- Traditional PDF viewers provide limited semantic search.
- General AI systems may lack access to private documents.
- Risk of unsupported AI-generated answers.
- Some systems provide limited source citations.
- Scanned PDFs may require separate OCR tools.
- Difficult to trace answers back to exact pages.
- Manual document comparison requires additional effort.

Proposed System

The proposed **Chat with Your PDF — Multi-Document Q&A with Source Citations** provides a centralized conversational platform for interacting with multiple PDF documents. Users can upload one or more PDFs through the application interface, and the document-processing layer extracts text, page numbers, headings, and supported metadata. OCR can be applied to scanned documents when required. The extracted content is cleaned and divided into meaningful chunks while preserving their original document and page references. An embedding model converts these chunks into vector representations and stores them in a vector database. When a user asks a question, the query-processing module converts the question into a semantic representation. The retrieval engine searches across all selected documents and identifies relevant passages based on semantic similarity. A reranking module can prioritize the most useful passages and remove less relevant context. The Generative AI engine creates a concise answer grounded in the retrieved content. The citation layer attaches document names, page numbers, sections, or source snippets to the generated response. Users can ask follow-up questions while maintaining the conversation context. The system can also provide an indication when sufficient evidence is not found rather than inventing an answer. Overall, the proposed system improves document accessibility, semantic search, multi-document analysis, and answer verification.

Advantages of Proposed System

- Supports multiple PDF documents.
- Provides natural-language document interaction.

- Uses semantic rather than only keyword-based retrieval.
- Searches across multiple documents simultaneously.
- Provides source citations with answers.
- Helps users verify generated information.
- Reduces time spent manually reading documents.
- Supports follow-up conversational questions.
- Can support scanned PDFs through OCR.
- Uses RAG to ground AI responses in retrieved content.
- Supports document-level and page-level references.
- Provides a centralized document-question-answering interface.

IV. METHODOLOGY

The methodology begins when users upload one or more PDF documents through the application interface. The document-processing module identifies the PDF structure and extracts available text, page numbers, headings, tables, and metadata. If a document contains scanned pages, an OCR module can convert the visible text into machine-readable content. The extracted content is cleaned by removing unnecessary formatting, repeated elements, and processing noise. The document is then divided into meaningful chunks while preserving the original document name and page references. An embedding model converts each chunk into a numerical vector representation. These vectors and their associated metadata are stored in a vector database for semantic retrieval. When a user submits a question, the query-processing module generates a semantic representation of the question. The retrieval engine searches the vector database across the selected documents and returns relevant content. A reranking stage can reorder retrieved passages based on relevance before they are provided to the Generative AI model. The AI model generates an answer using the retrieved context and is instructed not to rely on unsupported information. A citation module connects the generated answer to document names, page numbers, and relevant source passages. Finally, the response is displayed through the conversational interface, allowing the user to verify sources and ask additional questions.

SYSTEM ARCHITECTURE

The Chat with Your PDF architecture consists of several layers for document ingestion, processing, semantic retrieval, Generative AI, and source citation. The User Interface Layer allows users to upload multiple PDF files and ask questions through a conversational chat interface. The Document Ingestion Layer receives the uploaded files and validates their format and size. The PDF Processing Layer extracts text, page information, headings, tables, and metadata from the documents. The OCR Layer optionally processes scanned PDF pages and converts images into searchable text. The Text Processing Layer cleans the extracted information and divides it into meaningful chunks while preserving document and page references. The Embedding Layer converts chunks into vector representations using an embedding model. The Vector Database stores embeddings together with document metadata for efficient semantic search. The Query Processing Layer converts user questions into searchable semantic representations. The Retrieval and Reranking Layer retrieves and prioritizes relevant passages across multiple documents. The Generative AI Layer uses the retrieved context to generate a grounded natural-language answer. The Citation and Validation Layer connects the response to document names, page numbers, sections, or source

snippets and checks whether sufficient evidence exists. Finally, the Chat and Document Viewer Layer displays the answer and citations while allowing users to inspect the original document content.



V. RESULT AND OUTPUT

Chat with Your PDF
 Multi-Document Q&A with Source Citations

Home | Chat | Documents | Settings

“Turn your documents into answers.”

SR Srianjitha

My Documents (3)

- AI_in_Education.pdf (120 pages)
- Machine_Learning_Basics.pdf (85 pages)
- Cybersecurity_Guide.pdf (60 pages)

Chat History

- Main findings from documents (Today, 10:24 AM)
- Compare ML and cybersecurity (Today, 09:15 AM)
- Summary of education trends (Yesterday, 06:30 PM)
- Explain zero trust security (Yesterday, 04:12 PM)

Chat Conversation:

SR: What are the main findings from these documents?

10:24 AM

Based on the information from the uploaded documents, here are the main findings:

- AI in Education:** The document highlights that AI can personalize learning, automate administrative tasks, and improve student outcomes. It also emphasizes the importance of ethical use, data privacy, and the need for teacher training to effectively integrate AI tools in classrooms [1].
- Machine Learning Basics:** The document explains fundamental ML concepts such as supervised and unsupervised learning, common algorithms (e.g., linear regression, decision trees, neural networks), and real-world applications. It notes that the choice of algorithm depends on the problem type and available data [2].
- Cybersecurity Guide:** The document outlines key cybersecurity practices, including risk assessment, threat detection, access control, and the zero trust security model. It also discusses emerging threats and the importance of continuous monitoring and employee awareness [3].

Overall, the documents suggest that AI and machine learning are powerful tools with transformative potential, but they must be implemented responsibly with strong security and ethical considerations.

Sources

- 1. AI_in_Education.pdf (Page 12): “AI can personalize learning experiences and automate administrative tasks...”
- 2. Machine_Learning_Basics.pdf (Page 8): “Supervised learning uses labeled data to train models such as linear regression...”
- 3. Cybersecurity_Guide.pdf (Page 25): “The zero trust model assumes no user or device is trusted by default...”

Ask a question about your documents...

Summarize each document | Compare key concepts | Find information about implementation | What are the challenges?

Document Viewer:

AI_in_Education.pdf (Page 12 of 120)

3.2 Benefits of AI in Education

Artificial Intelligence (AI) has the potential to transform education by enabling personalized learning experiences, automating administrative tasks, and improving student outcomes. AI-powered tools can analyze student performance and provide tailored recommendations, helping learners progress at their own pace.

Furthermore, AI can assist educators in identifying students who may need additional support and suggest appropriate interventions. However, the successful implementation of AI in education requires careful planning, teacher training, and attention to ethical considerations such as data privacy and fairness.

12

Summary:

- 3 Documents Uploaded
- 12 Questions Asked
- ~5s Average Response Time
- 100% Answers with Citations

“Knowledge across documents. Answers with evidence.”

VI. CONCLUSION

The Chat with Your PDF — Multi-Document Q&A with Source Citations system provides an intelligent and convenient approach for interacting with large collections of PDF documents. Instead of manually searching through multiple files, users can upload their documents and ask questions using natural language through a single conversational interface.

The system combines PDF processing, OCR, Natural Language Processing, embeddings, vector databases, Retrieval-Augmented Generation (RAG), and Generative AI to retrieve relevant information and generate context-aware answers. The multi-document retrieval capability allows the system to identify useful information from several PDFs and combine the retrieved context into a meaningful response.

A key feature of the system is its source citation mechanism. The generated answers can include document names, page numbers, sections, and relevant source snippets, allowing users to verify the information against the original documents. This improves transparency and reduces reliance on unsupported AI-generated responses.

Overall, the proposed system reduces the time and effort required to study large documents and makes document analysis more interactive and accessible. It can be useful for students, researchers, professionals, organizations, legal teams, and analysts who frequently work with large PDF collections.

Future enhancements can include multimodal PDF understanding, improved OCR, table and image analysis, multilingual question answering, document comparison, advanced citation verification, voice-based interaction, enterprise access control, and agentic document retrieval. The system should continue to prioritize source grounding and clearly indicate when the uploaded documents do not contain sufficient information to answer a question.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int.

Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.

[12] V. N. S. Manaswini, K. K. C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in *Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT)*, Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[13] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[14] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[15] M. M. Rao, M. B. Mohiuddin, P. Srinu, A. Satyanarayana, J. Madhavan and R. D. Kumar, "Carbon-Neutral-Agriculture: Renewable-Energy Powered IoT Systems and Resource Optimization," *2026 3rd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, Chennai, India, 2026, pp. 1-6, doi: 10.1109/RMKMATE69073.2026.11518903.

[16] R. Uma, M. Maggidi, A. R. Ekkati, S. Bandlamudi, C. Madugundi and S. K. Taduri, "Zero-Trust Architecture for Hybrid Cloud Security in Critical Infrastructures," *2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON)*, Lonavala, India, 2026, pp. 1-6, doi: 10.1109/I3CTCON68242.2026.11507458.

- [17] VNS Manaswini, Chithanoor Arjun, Yamini Chouhan, & V. Sumalatha. (2026). MB-EFF: A Multi-Branch Environmental Feature Fusion Deep Learning Model for Crop Classification and Recommendation. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(22s), 460–468. <https://doi.org/10.70917/ijcisim-2026-5453>
- [18] Yamini Chouhan, SIDDOJU DIVYA, VNS Manaswini, & J. Priyanka. (2026). Vision-Threatening Diabetic Retinopathy Detection Using EfficientNetB4: A Two-Stage Transfer Learning Approach. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(22s), 469–476. <https://doi.org/10.70917/ijcisim-2026-5454>
- [19] Real Time Vehicle Counting and Classification system, **Dr. A.Satyanarayana,et.al** International Research Journal of Modernization in Engineering Technology and Science, Vol-06, Issue-05, 2024 [ISSN: 2582-5208] doi: 10.56726/IRJMETs57804
- [20] Facial Depth Maps: Detecting Sleep Apnea with Deep Learning, **Dr. A.Satyanarayana,et.al** , International Journal of Scientific Development and Research, Vol 9, Issue-5, 2024 [ISSN: 2455-2631]
- [21] Moodchef: Mood Based Smart Recipe Recommender with cooking Assistance, **Dr. A.Satyanarayana,et.al** , International Journal of Progressive Research in Engineering Management and Sciences, Vol-05,Issue-7, July 2025 [ISSN: 2583-5162] doi: 10.58257/IJPREMS42696
- [22] News Article Text Summarization, **Dr. A.Satyanarayana,et.al**, Journal of Emerging Technologies and Innovative Research, Vol-11, Issue -5, 2024, [ISSN: 2349-5162]
- [23] Voice Assistant and gesture controlled virtual mouse using Deep learning Techniques, **Dr. A.Satyanarayana,et.al** , International Research Journal of Modernization in Engineering Technology and Science Vol-6, Issue-5, 2024 [ISSN: 2582-5208] 10.56726/IRJMETs57214