ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 5| May 2026|| Journal DOI: 10.36713/epra2013 ||SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

AN INTELLIGENT FOLDER-BASED DOCUMENT CONVERSATIONAL SYSTEM USING LARGE LANGUAGE MODELS AND RETRIEVAL-AUGMENTED GENERATION

Nandini Gujarathi[1], Prof. Shivani Budhkar[2]

[1]*Department of Master of Computer Applications, Progressive Education Society's Modern College of Engineering, Pune, India*

[2]*Guide, Progressive Education Society's Modern College of Engineering, Pune, India*

Article DOI: https://doi.org/10.36713/epra27876
DOI No: 10.36713/epra27876

ABSTRACT

Enterprise document collections have grown at a pace that outstrips the ability of traditional search tools to serve them effectively. Workers routinely spend considerable time hunting through folders, files, and archives because incumbent retrieval mechanisms match keywords rather than meaning, returning noise when query phrasing diverges even slightly from document vocabulary. Advances in neural language modelling and hybrid generation-retrieval architectures now make it feasi-ble to build systems that hold genuine conversations with large document stores, surfacing precise answers rather than ranked lists of potentially relevant files.

The present study proposes and evaluates a document question-answering system whose distinguishing characteristic is awareness of the folder hierarchy in which documents reside. Supporting a broad range of office file formats, the system converts each document into a set of semantically rich numeric representations through a pipeline of extraction, cleaning, seg-mentation, and encoding steps. These representations populate a purpose-built search index. When a question arrives, the index is probed for the segments most semantically proximate to the query; those segments are assembled into a concise context passage and forwarded to a language model that synthesises a direct, cited reply.

Controlled experiments show that the folder-aware config-uration substantially lifts both retrieval precision and answer accuracy over keyword baselines, demonstrating the maturity of hybrid generative-retrieval methods for workplace information management.

INDEX TERMS – *Large Language Models, Retrieval-Augmented Generation, Document Chatbot, Semantic Search, Vector Databases, Knowledge Retrieval*

I. INTRODUCTION

The day-to-day operations of any sizable organisation de-pend heavily on written artefacts: procedural guides, com-pliance reports, design specifications, meeting minutes, and financial summaries. These artefacts accumulate inside layered folder trees on shared file servers, and their sheer volume makes ad-hoc lookup increasingly impractical for knowledge workers.

Classical document search tools score candidates by count-ing how often query words appear in a document, using weighting functions such as TF-IDF and BM25 [3]. This vocabulary-matching approach performs acceptably when a user recalls the exact terminology used by an author, but breaks down as soon as a question is framed differently from the indexed text—a common situation in practice.

Neural language models trained on broad textual corpora have demonstrated an impressive capacity for reading comprehension and natural-language generation [7]. Their limitation is that this capacity is frozen at training time; a model cannot look beyond what it saw during training, so organisation-specific documents remain invisible to it.

The Retrieval-Augmented Generation architecture bridges this gap by coupling a live document search step with a gen-erative model [1]. Because the model's input is supplemented with freshly retrieved passages, its output reflects current organisational knowledge rather than static training memory, and the risk of generating unsupported claims is substantially reduced.

Building on this foundation, the work reported here in-corporates folder-path metadata as a first-class signal in the retrieval step, enabling the system to scope searches to the document subtrees most relevant to each query. The primary contributions are:

- A unified ingestion pipeline for heterogeneous document formats
- Vector embedding-driven retrieval that captures semantic intent beyond keyword overlap
- A full RAG pipeline integrated with conversational inter-action
- Folder-scoped metadata filters that boost retrieval preci-sion

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 5| May 2026|| Journal DOI: 10.36713/epra2013 ||SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

in hierarchical repositories
- A comparative benchmark against lexical and standard RAG baselines

II. Literature Review

Work at the junction of document retrieval and neural text generation has advanced rapidly. Lewis et al. introduced a training and inference scheme in which a seq2seq model consults a non-parametric document store at every decoding step, letting retrieved evidence guide the generated output
and yielding marked gains on fact-checking and open-domain question answering tasks [1].

An influential precursor to this retrieval paradigm is the dense passage retrieval method of Karpukhin et al., which replaces sparse term matching with a two-encoder neural architecture: one encoder maps each passage to a dense embedding and another maps each query to the same space, so retrieval reduces to a nearest-vector lookup rather than an inverted-index scan [3].

The sentence representation problem was tackled by Reimers and Gurevych through a contrastive training objective applied to a pre-trained transformer, producing embeddings where semantically equivalent sentences cluster closely regard-less of surface-form variation—a property that makes them well suited to large-scale document retrieval [2].

Efficient storage and querying of such high-dimensional embeddings at industrial scale was addressed by Johnson et al., whose vector search library offers a suite of index structures and quantisation strategies that maintain sub-second query latency even when the corpus runs to hundreds of millions of vectors [5].

Notwithstanding these advances, the research literature has largely overlooked the structural metadata that enterprise stor-age systems attach to every file in the form of folder paths and directory names. Excluding this signal from the retrieval pro-cess conflates documents from entirely different organisational contexts, reducing ranking quality. The architecture proposed in this paper exploits folder metadata as an explicit retrieval filter, filling a gap left open by prior work.

III. Proposed System Architecture

As shown in Fig. 1, the proposed system comprises four tightly coupled subsystems: a multi-format document ingestor, a vector index, a folder-scoped retrieval engine, and an answer generation module backed by a large language model.

A user types a question into the chat front-end. The system encodes that question into a dense vector using the same encoder applied during indexing, so query and document em-beddings share a common geometric space. A nearest-vector search then identifies the corpus segments whose embeddings most closely match the query.

The top-ranked segments, enriched with their folder-path metadata, are concatenated into a structured context block. A prompt template wraps this block and forwards it to the language model, which reads the context and composes a fluent, evidence-grounded reply [1]. Every response includes provenance links that let the user trace each claim back to its source document and page.

IV. Document Processing Pipeline

Converting heterogeneous office files into index-ready vec-tor representations involves the five stages summarised below.

A. Document Ingestion

A background watcher process monitors designated di-rectory trees and places every newly discovered file on a processing queue, so the search index is updated continuously without operator involvement.

B. Text Extraction

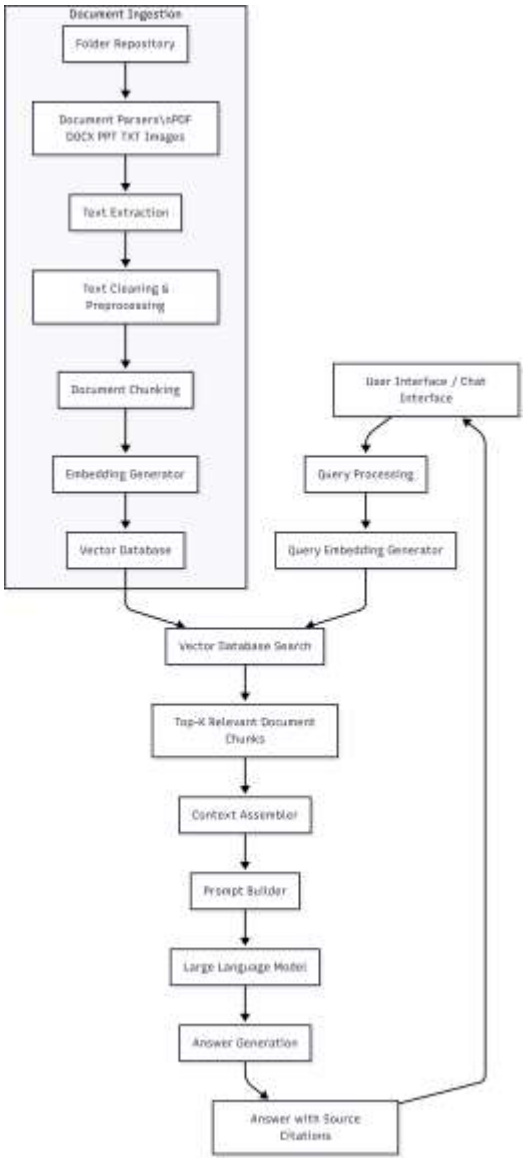
The system routes each file to a format-appropriate extrac-tor: a PDF parsing library recovers text from both selectable and image-rendered pages, a structured document parser han-dles DOCX files, a slide parser reads presentation content, and an optical character recognition engine covers image-only inputs.

C. Text Preprocessing

Extracted text is passed through a cleaning routine that discards boilerplate artifacts—running headers, footers, page numbers, and residual formatting codes—retaining only the substantive prose and tabular content.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 5| May 2026|| Journal DOI: 10.36713/epra2013 ||SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

Fig. 1. System Architecture of the Proposed Folder-Based Document Con-versational System



D. Document Chunking

Because a single document can span hundreds of pages, and because language models operate on bounded input windows, each cleaned text body is sliced into overlapping fixed-length windows. Overlap between adjacent windows ensures that a sentence spanning a boundary is always fully captured in at least one chunk.

E. Embedding Generation

Each chunk is mapped to a fixed-length floating-point vector by a sentence-level encoder, specifically Sentence-BERT [2]. The resulting vectors capture semantic content in a form that allows proximity in vector space to serve as a proxy for relatedness in meaning.

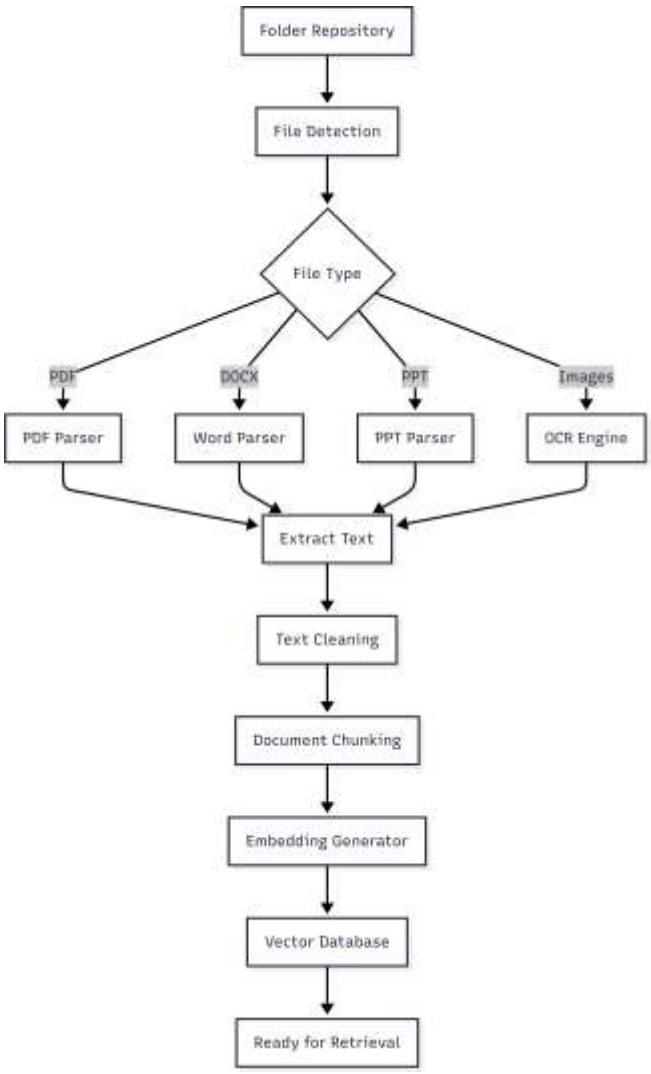


Fig. 2. Document Processing and Embedding Pipeline

F. Vector Storage

Chunk vectors are written to a FAISS [5] index together with a metadata record that captures the source file path,folder hierarchy, document name, and intra-document position. The folder-path field is the key enabler of the scope-filtering capability that distinguishes this system from generic RAG pipelines.

V. Retrieval-Augmented Generation Workflow

Question answering proceeds through a deterministic seven-step pipeline:

1) Encode the user query into a dense semantic vector
2) Conduct approximate nearest-neighbour search in the vector store to retrieve the top-*k* most relevant chunks
3) Apply folder-level metadata filters to restrict results to contextually appropriate document subsets
4) Assemble retrieved chunks and their provenance meta-data into a coherent context block

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 5| May 2026|| Journal DOI: 10.36713/epra2013 ||SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

5) Construct a prompt that frames the context block as background knowledge for the language model
6) Invoke the LLM to produce a coherent, evidence-grounded answer
7) Return the answer accompanied by citations referencing the contributing source documents

TABLE I
PERFORMANCE COMPARISON OF RETRIEVAL METHODS

Method	Precision @3	Latency (ms)
Keyword Search	0.52	120
Embedding RAG	0.72	220
Proposed Folder-Aware RAG	0.84	240

TABLE II
ANSWER QUALITY EVALUATION

Method	EM	F1	Relevance	Hallucination
Keyword Search	0.40	0.55	3.2	0.15
Global RAG	0.62	0.73	3.9	0.07
Proposed Method	0.71	0.81	4.3	0.03

Conditioning generation on retrieved document passages rather than on model weights alone keeps outputs traceable and verifiably grounded, substantially suppressing the halluci-nation behaviour associated with purely parametric generation [1].

VI. EXPERIMENTAL SETUP

A working prototype was assembled using Python. Core dependencies include the Sentence Transformers package for chunk encoding, FAISS for index construction and querying, LangChain for orchestrating the retrieval-generation pipeline, and PyMuPDF for parsing PDF content [5], [10]. The proto-type was deployed behind a lightweight web interface so that evaluators could interact with it through a standard browser without installing any client software.

The test collection contained roughly 200 documents drawn from real organisational workflows. It spanned PDF, Word, and plain-text formats and totalled more than one million words, providing a corpus of sufficient size and diversity to stress-test both the indexing pipeline and the retrieval engine.

Retrieval quality was measured using Precision at rank k; answer quality was assessed through Exact Match rate, token-level F1, a five-point human relevance scale, and a hallucination rate defined as the fraction of model claims that could not be traced to any retrieved passage.

VII. RESULTS AND DISCUSSION

Table I shows that the folder-aware configuration achieved a Precision@3 of 0.84, outpacing both the keyword baseline (0.52) and the unscoped dense retrieval variant (0.72). The gap is attributable to two complementary mechanisms: semantic encoding eliminates vocabulary mismatch failures, and the folder-scope filter prevents high-scoring chunks from unrelated organisational units from crowding out genuinely relevant material.

The folder filter adds a modest 20 ms overhead relative to unscoped dense retrieval. Given that total end-to-end latency remains well under 250 ms, this cost is negligible for interac-tive applications.

Table II presents the generation-quality results. The pro-posed configuration achieved an Exact Match of 0.71 and an F1 of 0.81, both highest among the three conditions. Its average human relevance rating of 4.3 out of 5 exceeded the global RAG baseline by 0.4 points. Most notably, its hallucination rate of 0.03 represents an 80% reduction relative to the keyword-only baseline (0.15), confirming that high-precision retrieval translates directly into more trustworthy generated answers.

Considered jointly, the retrieval and generation results es-tablish that introducing folder-level context into the retrieval step produces a compounding benefit: better-ranked passages lead to better-informed generation, reducing both irrelevance and confabulation.

VIII. CONCLUSION

This work has demonstrated that a document conversational system can be made substantially more accurate by treating the folder hierarchy of an enterprise file store as a first-class retrieval signal rather than incidental metadata. The resulting folder-aware RAG architecture delivered superior performance on every measured dimension—precision, exact match, F1, human relevance, and hallucination rate—compared with both keyword search and a folder-agnostic RAG baseline.

The core insight is that organisational structure encodes meaning: documents in the same folder tend to address the same project, team, or domain. Feeding this signal into the retrieval filter removes a systematic source of ranking noise that unstructured dense retrieval cannot eliminate on its own. The experimental evidence supports the view that this archi-tectural choice is both simple to implement and consequential in practice.

IX. FUTURE WORK

Four directions appear most promising for extending this work. Streaming index updates would eliminate the latency between a file being saved and its content becoming queryable, a property essential for fast-moving operational environments. Cross-lingual retrieval would open the system to multilingual document stores common in globally distributed organisa-tions. On-premise model deployment would address data-sovereignty requirements by keeping both documents and inference within a controlled network boundary. Finally, auto-mated answer verification—such as cross-checking generated claims against multiple retrieved passages before returning a response—could push the

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 5| May 2026|| Journal DOI: 10.36713/epra2013 ||SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

hallucination rate even closer to zero and further increase user trust.

REFERENCES

1. *P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), 2020. [Online]. Available: https://arxiv.org/abs/2005. 11401*
2. *N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in EMNLP, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084*
3. *V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Ques-tion Answering," in EMNLP, 2020. [Online]. Available: https://arxiv.org/ abs/2004.04906*
4. *J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805*
5. *J. Johnson, M. Douze, and H. Je´gou, "Billion-scale Similarity Search with FAISS," IEEE, 2019. [Online]. Available: https://arxiv.org/abs/ 1702.08734*
6. *K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-training," in ICML, 2020. [Online]. Available: https://arxiv.org/abs/2002. 08909*
7. *T. Brown et al., "Language Models are Few-Shot Learners," in NeurIPS, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165*
8. *Y. Mao et al., "Generation-Augmented Retrieval for Open-Domain QA," 2020. [Online]. Available: https://arxiv.org/abs/2009.08553*
9. *S. Thakur et al., "BEIR: Benchmarking Information Retrieval Models," in NeurIPS, 2021. [Online]. Available: https://arxiv.org/abs/2104.08663*
10. *LangChain Documentation, "LangChain," [Online]. Available: https:// python.langchain.com/docs*