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 4, November 2026
Impact Factor: 8.2

Project Spear: An AI-Powered Digital Memory Assistant Using Retrieval-Augmented Generation with Secure Access Control

Muhammed Suhail TS[1], Muhammed Shammas[2], Muhammed Shahid KM[3], Thanseeh PP[4], Niveditha P S[5]
Department of Computer Science and Engineering[1-5]
Thejus Engineering College, Thrissur, Kerala, India
Affiliated to APJ Abdul Kalam Technological University (KTU)

Abstract : *Information kept in personal and organisational documents is scattered across devices and platforms, and conventional keyword search and general-purpose chatbots cannot answer questions about it reliably. This paper presents Project Spear, an AI-powered digital memory assistant that answers natural language questions from a user's own documents using Retrieval-Augmented Generation (RAG). The proposed framework works in two phases. In offline indexing, uploaded documents are loaded, preprocessed, split into chunks, converted into embeddings and stored in a vector database (FAISS or ChromaDB). In the online query phase, the question is embedded, the most similar chunks are retrieved and optionally reranked, combined with the query into a context, and passed to a large language model that generates a grounded answer with optional citations. Authentication and access control, a React.js frontend, a FastAPI/Flask backend and a PostgreSQL/SQLite store for users and metadata complete the system. A survey of twenty recent RAG studies on evaluation, long-context handling, retrieval reliability, and security and privacy shows that none offers a complete, secure assistant for personal digital memory, which motivates the design. This paper reports the design phase of the project (architecture, data flow and requirements); experimental results are not yet reported.*

Keywords: Retrieval-Augmented Generation, Digital Memory, Vector Database, Large Language Models, Document Question Answering, Privacy.

I. INTRODUCTION

Individuals, students and organisations now keep a large share of their knowledge in documents such as notes, reports and study material. These documents end up spread across many devices and platforms, so finding one fact usually means remembering where it was saved and searching each place by hand. The tools in common use do not remove this effort. Keyword-based search matches only the words typed by the user and cannot understand the context of a question, and it has no memory of what a particular user has stored or asked before.

Large Language Models (LLMs) allow information to be queried in natural language, but a general-purpose chatbot has no access to a user's private documents. When asked about them it gives generic answers and may produce fabricated content, a problem known as hallucination. Retrieval-Augmented Genera-tion (RAG) addresses this by retrieving relevant passages from a document store and supplying them to the LLM as context, so the answer is grounded in real data. RAG-based systems have been proposed for intelligent document processing and study assistance [1], the RAG lifecycle has been examined for data spaces [2], retrieval, reranking and similarity choices have been evaluated empirically in question answering [3], and RAG for large language models has been surveyed [4].

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 4, November 2026 Impact Factor: 8.2

Despite this progress, a personal digital memory raises re-quirements that are rarely addressed together: answers must come from the user's own documents, new documents must be added at any time, and the stored data must be protected from other users and from leakage. In this work we present Project Spear, an AI-powered digital memory assistant that combines RAG, a vector database and access control. The main contribu-tions are summarised as follows:

- A two-phase RAG framework (offline indexing and online query) for answering natural language questions from a user's own documents, with optional reranking and citations.
- A secure application layer with authentication, access control and a relational store for users and metadata.
- A survey of twenty recent RAG studies, grouped into four themes, that identifies the lack of a complete, secure assistant for personal digital memory.
- A data-flow design (Level 0 and Level 1 DFDs) and a hard-ware and software specification for implementing the system.

The remainder of this paper is organised as follows. Section II reviews the literature, Section III presents the design of the system, and Section IV concludes the paper and outlines future work.

II. LITERATURE SURVEY

RAG has been studied extensively. Twenty recent studies were reviewed and grouped into four themes: evaluation of RAG, com-parison of RAG with long-context LLMs, reliability of retrieval, and security and privacy. Table I summarises the themes and their main limitations.

Evaluation of RAG. Es et al. [5] proposed RAGAs, a set of automated metrics for retrieval quality, response faithfulness and answer relevance. Yu et al. [6] surveyed existing RAG evaluation methods and the criteria used to judge retrieval and generated responses, and Qi et al. [7] introduced Long2RAG, which checks through Key Point Recall whether important in-formation is retrieved when documents and answers are long. Automated metrics may not match human judgment, the survey gives no implementation, and the long-context benchmark does not address secure storage of personal documents.

RAG versus long-context LLMs. Li et al. [8] compared RAG with long-context LLMs and proposed a hybrid approach, while LaRA [9] found no single approach best for every task. Jin et al. [10] showed that irrelevant retrieved information affects gen-eration quality in long-context LLMs, and hierarchical document refinement [11] reduces redundancy and cost in long-document RAG. Laitenberger et al. [12] asked whether complex RAG pipelines are needed at all and compared them with simpler baselines. These works add complexity or compute, depend on the chosen model and dataset, and are evaluated mainly on benchmarks rather than personal document collections.

TABLE I: SUMMARY OF THE LITERATURE SURVEY

Theme	Refs.	Main limitation
Evaluation of RAG	[5–7]	Metrics may differ from human judgment; no implementation; no secure storage
RAG vs. long-context LLMs	[8–12]	Added complexity and compute; results depend on model and dataset; benchmark-only
Retrieval reliability	[13–15]	Extra processing; overhead; ontology maintenance
Security and privacy	[16–24]	Threat analyses rather than a complete system; cost, complexity or lower accuracy

Reliability of retrieval. Astute RAG [13] handles irrelevant, conflicting, misleading or potentially malicious retrieved infor-mation. Li and Ouyang [14] studied how selecting information from retrieved knowledge affects response quality, and OG-RAG [15] brings ontology-based structures into retrieval for specialised domains. The extra processing and selection over-head raises complexity, and ontologies are hard to maintain for continuously changing personal information.

Security and privacy. Zeng et al. [16] examined how pri-vate or proprietary data can be exposed through retrieval, and SafeRAG [17] benchmarks security threats from manipulated or untrusted retrieved content. Defences include locally

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 4, November 2026
Impact Factor: 8.2

private entity perturbation [18], pure synthetic data [19], provably se-cure RAG against leakage and data poisoning [21], controlled retrieval with context minimisation [23], and storing document knowledge in model parameters instead of exposing the docu-ments [24]. Arzanipour et al. [20] formalised the RAG threat model and attack surface, and Palanisamy et al. [22] reviewed security and privacy across the pipeline. These mechanisms can reduce retrieval accuracy, increase computation and complexity, or make it hard to update changing information.

Most studies address one aspect of RAG on benchmark data, and the security work concentrates on threat analysis rather than a working system. None provides a complete, secure, end-to-end assistant for managing personal documents, which is the gap that Project Spear addresses.

III. DESIGN

A. Overview

Project Spear is a web application. The user works through a React.js frontend that sends requests to a FastAPI/Flask backend. Requests pass through authentication and access control, and user and metadata records are stored in PostgreSQL or SQLite. Behind the backend, the RAG system operates in two phases,

B. Document Ingestion (Offline Indexing)

Documents from sources such as PDFs, web pages and databases are collected by the document loader. Preprocessing extracts and cleans the text, and chunking splits it into smaller pieces so that each piece can be retrieved independently. The embedding model converts every chunk from text into a vector, and the vectors are stored in the vector database (FAISS or ChromaDB). Because indexing only adds new vectors, users can upload new documents at any time and the repository grows without retraining any model.

C. Query Processing and Retrieval (Online Query)

The user's question is converted into a vector by the same em-bedding model. The retriever then performs a similarity search in the vector database. If eq is the query embedding and ei the embedding of chunk ci, the similarity can be computed, for example, by cosine similarity:

$$s_i = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{\|\mathbf{e}_q\| \, \|\mathbf{e}_i\|} \qquad (1)$$

The top-K chunks with the largest scores form the retrieved set
R, and optional reranking can reorder them before use:

$$\mathscr{R} = \{c_i : s_i \text{ is among the } K \text{ largest scores}\} \qquad (2)$$

D. Grounded Answer Generation

The retrieved chunks are combined with the query to build the context C , and the LLM generator produces the answer A from this context:

$$\mathscr{C} = \text{Build}(\mathscr{R}, q), \qquad A = \text{LLM}(\mathscr{C}) \qquad (3)$$

The same mechanism produces summaries of documents. The final response may include citations that point back to the source chunks. Because the LLM answers from the retrieved context rather than from its own memory alone, hallucinations are re-duced.

E. Secure Access and Metadata Management

Authentication and access control ensure that each user can reach only their own documents and metadata. User accounts and document metadata are stored in PostgreSQL or SQLite, while embeddings are kept in the vector

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 4, November 2026
Impact Factor: 8.2

database. This separation keeps user management independent of retrieval and responds to the privacy concerns identified in the survey.

F. Data Flow Design

The Level 0 DFD (Fig. 2a) shows the user as the only external entity. The user supplies credentials, documents and questions to the system, which is a single process, and receives grounded answers and summaries. The Level 1 DFD (Fig. 2b) breaks the system into four processes and two data stores:

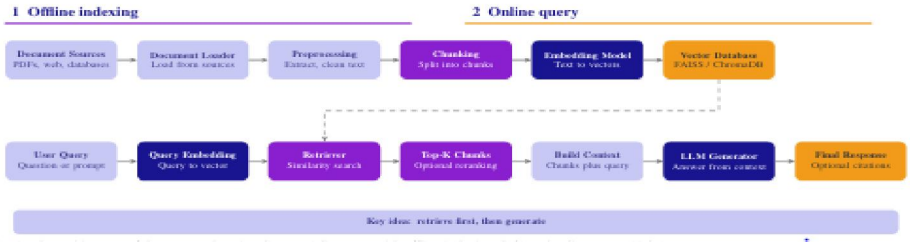
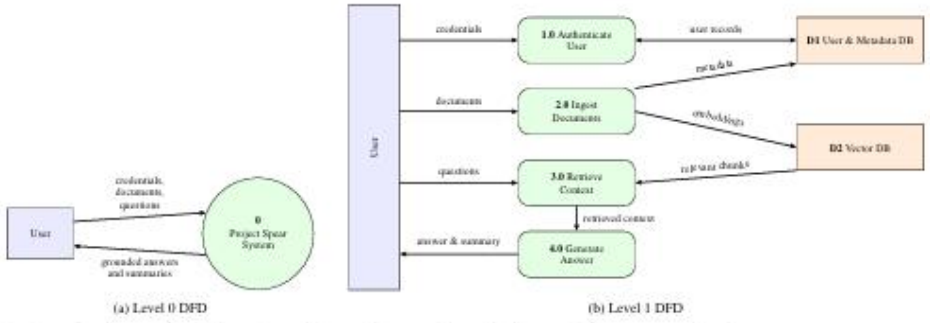


Fig. 1. The architecture of the proposed Project Spear RAG system with offline indexing (left) and online query (right).



(a) Level 0 DFD (b) Level 1 DFD

Fig. 2. Data flow diagrams of Project Spear: (a) Level 0 context diagram and (b) Level 1 diagram with four processes and two data stores.

- **1.0 Authenticate User:** verifies credentials and applies access control using the user database (D1).
- **2.0 Ingest Documents:** extracts text, chunks it and gener-ates embeddings stored in the vector database (D2), with document metadata kept in D1.
- **3.0 Retrieve Context:** embeds the question and retrieves the relevant chunks from D2.
- **4.0 Generate Answer**: the LLM generates a grounded answer and summary from the retrieved context and returns it to the user.

G. System Configuration

Project Spear performs document processing, embedding gener-ation and retrieval, which need a moderately powerful machine. The configuration is given in Table II. A mid-range multi-core processor and 8 GB of RAM (16 GB recommended) are needed for the backend, text extraction, chunking and embedding gener-ation. An SSD speeds up reading and writing during indexing and retrieval, and a GPU is optional but can speed up embedding generation and model inference when models are run locally. SQLite is lightweight for development, whereas PostgreSQL suits multi-user use.

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 4, November 2026
Impact Factor: 8.2

TABLE II: SYSTEM CONFIGURATION OF PROJECT SPEAR

Parameter	Value
Operating system	Windows 10/11 or Ubuntu Linux
Programming language	Python 3.11+
Frontend	React.js / HTML, CSS, JavaScript
Backend	FastAPI or Flask
Metadata database	PostgreSQL / SQLite
Vector database	FAISS or ChromaDB
Processor	Intel Core i5 / AMD Ryzen 5 or above
RAM	8 GB (16 GB recommended)
Storage	256 GB SSD (512 GB recommended)
GPU	NVIDIA RTX 3050/4050 (optional)
Internet	Stable broadband connection
Development tools	Visual Studio Code; Git and GitHub

H. Comparison with Existing Approaches

Table III compares keyword-based search, general-purpose chat-bots and the proposed system on the problems identified in this work. Keyword search lacks context understanding and relies only on word matching, so it returns files rather than answers. General chatbots understand natural language but have no per-sonalised memory, cannot retrieve user-specific information and are prone to hallucination. Project Spear is designed to address

TABLE III: QUALITATIVE COMPARISON OF APPROACHES

Feature	Keyword search	General chatbot	Project Spear
Context understanding	No	Yes	Yes
Uses the user's own documents	Word match only	No	Yes
Personalised memory	No	No	Yes
Documents in one place	No	No	Yes
Answers grounded in retrieved text	—	No	Yes
Authentication and access control	—	—	Yes (designed)

these limitations by grounding answers in a centralised, per-user document repository.

IV. CONCLUSION AND FUTURE WORK

This work presented Project Spear, an AI-powered digital mem-ory assistant that stores a user's documents in one place and answers natural language questions from them using Retrieval-Augmented Generation. The framework combines offline in-dexing with a vector database, online retrieval with optional reranking, grounded answer generation with optional citations, and authentication and access control. A survey of twenty recent studies showed that no existing work offers a complete, secure assistant for personal digital memory. By grounding answers in retrieved documents, the design aims to reduce hallucinations and manual search time while remaining scalable and secure. It is expected to give accurate, context-aware responses, to support multiple document formats, and to stay up to date because users can upload new documents at any time without retraining.

This paper presents the design phase only. The system has not yet been evaluated, so no accuracy, latency or user-study results are reported, and the benefits above are design expectations rather than measured findings. Answer quality will depend on retrieval and chunking quality, and privacy mechanisms from the survey, such as entity perturbation or controlled retrieval, may trade accuracy for protection.

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 4, November 2026 Impact Factor: 8.2

Future work will implement and evaluate the system, for ex-ample with automated metrics such as RAGAs [5], and extend it with support for images, scanned documents and tables [1]; mem-ory of previous conversations; improved retrieval through rerank-ing and knowledge selection [3, 14]; stronger privacy through controlled retrieval and sensitive-entity protection [18, 23]; han-dling of very long documents by combining RAG with long-context LLMs [8, 9]; multi-user enterprise deployment with role-based access control; and a mobile application with cloud storage integration.

REFERENCES

1. H. Kunder, C. M. Darshan, S. S. Shetty, T. Sathyaprakash, and S. S. Poo-jary, "A multimodal retrieval-augmented generation system for intelligent document processing and study assistance," in Proc. 2026 Int. Conf. Artifi-cial Intelligence and Data Engineering (AIDE), Nitte, India, Feb. 2026, pp. 416–421, doi: 10.1109/AIDE69088.2026.11544235.
2. M. Alqatf, R. Haque, S. H. Alsamhi, S. Buosi, M. A. Razzaq, M. Timilsina, Hawbani, and E. Curry, "RAG4DS: Retrieval-augmented generation for data spaces—A unified lifecycle, challenges, and opportunities," IEEE Access, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3545387.
3. H. Elkiran and J. Rasheed, "An empirical evaluation of retrieval, rerank-ing, and similarity for a Q&A-based retrieval augmented generation sys-tem," IEEE Access, vol. 14, pp. 26053–26066, 2026, doi: 10.1109/AC-CESS.2026.3664852.
4. "A survey of retrieval-augmented generation (RAG) for large language models," in 2025 Int. Conf. Trustworthy Big Data and Artificial Intelligence (ICTBAI), Aug. 21–23, 2025, doi: 10.1109/ICTBAI68361.2025.00008.
5. S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," 2024.
6. H. Yu et al., "Evaluation of retrieval-augmented generation: A survey," 2024.
7. Z. Qi et al., "Long2RAG: Evaluating long-context & long-form retrieval-augmented generation with key point recall," 2024.
8. Z. Li et al., "Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach," 2024.
9. K. Li et al., "LaRA: Benchmarking retrieval-augmented generation and long-context LLMs—No silver bullet for LC or RAG routing," 2025.
10. B. Jin, J. Yoon, J. Han, and S. Arik, "Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG," 2025.
11. J. Jin et al., "Hierarchical document refinement for long-context retrieval-augmented generation," 2025.
12. A. Laitenberger, C. D. Manning, and N. F. Liu, "Stronger baselines for retrieval-augmented generation with long-context language models," 2025.
13. F. Wang et al., "Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models," 2025.
14. X. Li and J. Ouyang, "How does knowledge selection help retrieval aug-mented generation?," 2025.
15. K. Sharma, P. Kumar, and Y. Li, "OG-RAG: Ontology-grounded retrieval-augmented generation for large language models," 2025.
16. S. Zeng et al., "The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)," 2024.
17. X. Liang et al., "SafeRAG: Benchmarking security in retrieval-augmented generation of large language model," 2025.
18. L. He et al., "Mitigating privacy risks in retrieval-augmented generation via locally private entity perturbation," 2025.

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 4, November 2026
Impact Factor: 8.2

19. S. Zeng et al., "Mitigating the privacy issues in retrieval-augmented gener-ation (RAG) via pure synthetic data," 2025.
20. A. Arzanipour, R. Behnia, R. Ebrahimi, and K. Dutta, "RAG security and privacy: Formalizing the threat model and attack surface," 2025.
21. P. Zhou, Y. Feng, and Z. Yang, "Provably secure retrieval-augmented generation," 2025.
22. B. Palanisamy, G. S. S. Chalapathi, V. Hassija, and R. Buyya, "Security and privacy in retrieval-augmented generation: Architectures, threats, defenses, and future directions for building trustworthy systems," 2026.
23. "Private-RAG: A privacy-preserving retrieval-augmented generation method for large model inference," 2026.
24. J. Chen et al., "Privacy-preserving reasoning with knowledge-distilled parametric retrieval augmented generation," 2026.