Journal of Library Insights & Innovations (JLII)
An open access peer reviewed scholarly publication dedicated to advancing the field of library & information science

Volume 1, Issue 2
December 2025

Reimagining Academic Library Services through Local RAG: Concepts and Frameworks

DOI: 10.63880/jlii.v1i2.49

Pawan Kumar Pal [1]

ABSTRACT

Purpose: *This study examines Local Retrieval-Augmented Generation as an emerging artificial intelligence framework for strengthening information services in academic libraries. It addresses the limitations of traditional keyword-based retrieval systems and cloud-based language models, particularly issues related to semantic accuracy, unreliable responses, data privacy, and institutional dependency. The study aims to conceptualize Local Retrieval-Augmented Generation for library contexts and to outline its potential role in modernizing academic information services while preserving institutional control.*

Methodology: *The study adopts a conceptual and analytical research design based on a critical review of recent literature and practical implementations related to artificial intelligence, language models, and retrieval systems in libraries. A structured conceptual architecture of Local Retrieval-Augmented Generation is developed, followed by the formulation of a six-phase deployment framework designed specifically for academic library environments.*

Findings: *The analysis indicates that Local Retrieval-Augmented Generation enhances information accessibility through conversational interfaces and improves retrieval accuracy by grounding responses in locally curated institutional documents. It supports complete data sovereignty, offers cost-effective deployment options, and allows high levels of system customization. However, implementation challenges include technical complexity, infrastructure and computational demands, the need for staff training, and ethical and governance considerations.*

Received:10.12.2025
Revised: 22.12.2025
Accepted: 28.12.2025
Published: 30.12.2025

[1] Research Scholar, DLIS, *University of Calcutta, Email- pawanpalcu@gmail.com*

Implications: *The study concludes that Local Retrieval-Augmented Generation is a viable and strategic solution for academic libraries seeking to modernize services while maintaining autonomy over data and systems. Successful adoption requires systematic planning, capacity building, and sustained institutional commitment to ethical and responsible artificial intelligence governance.*

Keywords: Retrieval-Augmented Generation, Local RAG, Artificial Intelligence, Open-source LLM, information retrieval, academic library, conversational AI, library technology

1. INTRODUCTION

Academic libraries play a critical role in supporting learning and research activities within higher education institutions by facilitating access to reliable and organized information resources (Khan et al., 2023). With the rapid expansion of digital collections and evolving user expectations, libraries are increasingly required to adopt intelligent technologies that can enhance information discovery and user engagement (Cox et al., 2019).

Traditional keyword-based retrieval systems rely on exact term matching and often fail to capture semantic relationships between concepts, thereby limiting retrieval precision and user satisfaction (Roy et al., 2012). These limitations have become more pronounced in digitally intensive research environments, where users expect interactive and personalized information services.

Recent advances in Large Language Models (LLMs) have introduced conversational interfaces that support natural language interaction for information access (Lewis et al., 2020). However, studies have highlighted that such models may generate hallucinated or inaccurate responses due to their reliance on fixed pre-trained datasets rather than institution-specific knowledge sources (Li et al., 2023). Additionally, the dependence of commercial LLMs on cloud-based infrastructures raises concerns related to data privacy, institutional governance and long-term sustainability within academic library environments (Wang et al., 2024).

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address these challenges by integrating document retrieval mechanisms with generative language models, thereby grounding responses in verifiable sources. While existing RAG implementations demonstrate improved response accuracy (Lewis et al., 2020; Gao et al., 2023), most remain dependent on cloud-based infrastructures, which perpetuate concerns regarding data privacy and vendor dependence.

In this context, Local RAG offers a locally deployed open-source alternative that may enable academic libraries to retain control over institutional data, customize systems to local collections and mitigate privacy risks associated with cloud-based solutions (Mukhopadhyay, 2025; Xuan, 2025). Although Retrieval-Augmented Generation has been widely discussed in technical and computational research, existing studies largely focus on cloud-based architectures and system-level performance. Within Library and Information Science literature, discussions on artificial intelligence primarily emphasize chatbots and generic AI applications, with limited attention to locally deployed RAG systems. In particular, there is a lack of structured, library-oriented conceptual frameworks that address implementation planning,

DOI: 10.63880/jlii.v1i2.49

institutional readiness, governance and sustainability. This study addresses this gap by presenting a comprehensive conceptual analysis of Local RAG and proposing a phased implementation framework tailored to academic library environments.

2. LITERATURE REVIEW

The development of information retrieval systems in academic libraries has progressed from traditional keyword-based search systems to intelligent and automated discovery platforms. Roy et al. (2012) explained that keyword-based retrieval depends primarily on exact term matching, which restricts the system's ability to recognize semantic relationships between concepts. Huang (2022) further observed that such limitations negatively affect retrieval accuracy, particularly in multidisciplinary and complex research environments.

Studies have also shown that the rapid growth of digital collections has intensified these challenges for academic libraries. Cox et al. (2019) noted that traditional retrieval systems provide limited personalization and contextual support for users. Similarly, Khan et al. (2023) emphasized that evolving user expectations now demand interactive and intelligent information services that go beyond traditional search interfaces.

Recent literature highlights the growing adoption of artificial intelligence technologies, particularly Large Language Models (LLM) to address these limitations. Li et al. (2023) demonstrated that LLMs may generate hallucinated or inaccurate responses due to their reliance on pre-trained data rather than real-time or institution-specific sources. Ji et al. (2022) further identified hallucination as a persistent challenge in natural language generation systems. Wang et al. (2024) highlighted that dependence on external AI service providers may conflict with institutional governance policies, particularly in sensitive academic and research environments.

To mitigate these issues, researchers have increasingly focused on Retrieval-Augmented Generation (RAG) as an alternative approach. Lewis et al. (2020) introduced RAG as a framework that integrates document retrieval with generative models to ensure responses are grounded in retrieved source documents. Gao et al. (2023) later expanded on this concept, noting that RAG improves factual consistency by combining retrieval and generation processes. Brown et al. (2025) observed that RAG-based systems significantly reduce hallucination compared to standalone language models. Bevara et al. (2025) further reported that RAG has strong potential for improving search and retrieval services within academic libraries.

Despite these advantages, existing studies indicate that most RAG implementations remain dependent on cloud-based infrastructures. Mukhopadhyay (2025) pointed out that such reliance continues to pose challenges related to data sovereignty and vendor dependence. Xuan (2025) emphasized that these concerns have encouraged interest in Local RAG models that operate entirely within institutional environments. Johnson et al. (2017) demonstrated that vector databases enable efficient semantic similarity search, which forms a core component of Local RAG systems. Reimers and Gurevych (2019) further contributed by introducing sentence embedding techniques that enhance semantic representation of textual content. However, existing literature largely concentrates on technical feasibility and system architecture, while

DOI: 10.63880/jlii.v1i2.49

limited attention is given to structured implementation strategies, organizational readiness and governance considerations specific to academic libraries.

Therefore, it is evident from the reviewed literature that a conceptual gap exists in systematically addressing Local RAG from a library-centered perspective. The present study addresses this gap by synthesizing existing research to develop a conceptual framework and practical guidance for implementing Local RAG in academic library services.

3. OBJECTIVES OF THE STUDY

The objectives of this article are as follows:

- To examine the conceptual framework and technical architecture of Local RAG systems.
- To identify advantages, opportunities and challenges of implementing Local RAG in academic libraries.
- To provide practical implementation frameworks and recommendations for academic library professionals regarding Local RAG deployment.

4. RESEARCH METHODOLOGY

This study adopts a systematic qualitative synthesis approach grounded in thematic analysis to examine Local RAG as an emerging artificial intelligence framework for academic library services. The research focuses on synthesizing existing literature to develop a structured conceptual understanding of RAG architecture, implementation prospects and associated challenges within academic library contexts.

A systematic literature review was conducted to identify relevant studies related to Retrieval-Augmented Generation, large language models, conversational AI and artificial intelligence applications in academic libraries. Scholarly sources were collected from recognized academic databases from Scopus, Google Scholar, ArXiv and ResearchGate. The review was limited to English-language publications published between 2020 and 2025 to ensure relevance to recent technological developments. Key search terms used during the literature retrieval process included *"retrieval-augmented generation," "Local RAG," "AI in academic libraries," "library chatbots,"* and *"local large language models."* From the initial set of retrieved records of 86 papers, 35 scholarly articles were selected based on their relevance to RAG system architecture, implementation practices and library-specific applications. The selected literature was analyzed using a thematic analysis approach, which involved identifying concepts, patterns and themes across the reviewed studies. The analysis focused on four major thematic areas: (i) evolution of information retrieval systems, (ii) architectural components of RAG and Local RAG systems, (iii) implementation opportunities and benefits for academic libraries, (iv) technical, organizational and ethical challenges associated with Local RAG adoption.

DOI: 10.63880/jlii.v1i2.49

Table 1. Summary of Thematic Analysis Identified from the Reviewed Literature (n = 35)

Thematic Area	Extent of Coverage	Key Aspects Examined	Focus of Analysis
i) Evolution of Information Retrieval Systems	High	Conceptual and historical analysis	Transition from traditional to intelligent systems
ii) RAG and Local RAG Architecture	Very High	Technical and architectural discussion	Structural components of RAG systems
iii) Implementation Opportunities in Academic Libraries	Moderate	Service-oriented applications	Service enhancement potential
iv) Challenges in Local RAG Adoption	Moderate	Organizational and ethical concerns	Constraints affecting implementation

This qualitative synthesis enabled the development of a conceptual framework and a practical implementation model grounded in existing literature. By organizing and interpreting current research findings thematically, the study provides structured insights that can inform empirical investigations and guide academic library professionals considering the adoption of Local RAG systems.

5. CONCEPTUAL FRAMEWORK OF LOCAL RAG

5.1 Architecture of Local RAG

RAG refers to an artificial intelligence framework that integrates the language model capabilities with document retrieval systems that are entirely deployed in an institutional context based on open-source systems (Lewis et al., 2020; Gao et al., 2023). Unlike standalone large language models that rely solely on pre-trained knowledge, Local RAG systems retrieve relevant documents from institutional repositories before generating responses, grounding outputs in verifiable sources rather than model-generated content (Brown et al., 2025). This method has a significant impact on decreasing hallucinations and allowing systems to operate with existing proprietary or special institutional data (Bevara et al., 2025; Oche et al., 2025).

Figure 1. shows that the Local RAG workflow is implemented by three integrated engines. The Retrieval Engine includes a query processor and a search engine that carry out the operation of the vector searches using the query as interpreted by the library user against the indexed knowledge base (Johnson et al., 2017). The Augmentation Engine subsequently develops augmented prompts by injecting the retrieved pertinent documents as context. Lastly the Generation Engine utilizes a local generative AI model to generate grounded library user responses with citations generated based on the augmented context as opposed to pre-trained knowledge by itself (Mazumder and Mukhopadhyay, 2024).

DOI: 10.63880/jlii.v1i2.49

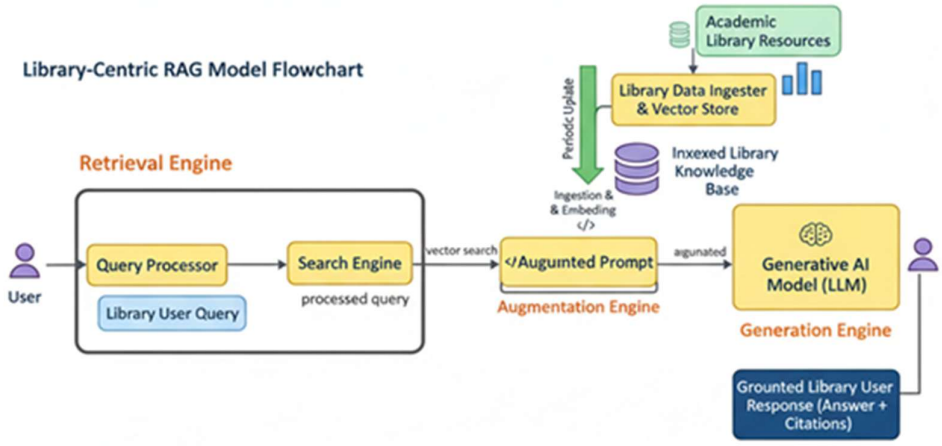

Figure 1. Library-centric Local RAG architecture showing the three-engine workflow

This workflow is facilitated by several major elements. Resources in academic libraries are updated periodically with the help of a library data ingester and a vector store converting institutional literature into comprehensible forms that are stored in an indexed body of knowledge (Reimers & Gurevych, 2019). Instead of utilizing cloud computing providers open-source language models like Mistral 7B or LLaMA can be executed directly within the facilities of an institution where the full data sovereignty is raised (Jiang et al., 2023).

The main distinction between Local RAG and cloud-based AI services is that there are no external data transmissions and the entire processing is under institutional control (Xuan, 2025). Contrary to the traditional library search systems based on the use of keywords matching Local RAG is semantically based and produces conversational replies. These are the features that make Local RAG especially appropriate to academic libraries that need the privacy of their data, the ability to customize it through institutional control and the customization options (Mukhopadhyay, 2025).

5.2 Open-Source Technology Stack for Local RAG Deployment

Several open-source tools enable academic libraries to implement Local RAG systems without commercial dependencies. Table 2. presents the essential technical components required for Local RAG architecture identifying available frameworks and their specific advantages for institutional library environments. These tools collectively support privacy-preserving customizable deployment within existing library infrastructure.

Table 2. Open-Source Tools and Technologies for Local RAG Implementation

Component	Tool/Framework	Primary Function	Key Features for Libraries
Language Model	Mistral 7B, LLaMA 2/3, Phi-3	Response generation	Multilingual support, customizable, runs on standard servers
Embedding Model	Sentence-BERT, all-MiniLM-L6-v2	Semantic text representation	Efficient similarity search, low computational cost

DOI: 10.63880/jlii.v1i2.49

Vector Database	ChromaDB, FAISS, Weaviate	Document storage and retrieval	Scalable for large collections, fast query response
RAG Framework	LangChain, LlamaIndex, Haystack	Workflow orchestration	Integration with existing systems, customizable pipelines
Local Deployment	Ollama, LocalAI, vLLM	On-premise LLM hosting	Privacy-preserving, no internet dependency
Document Processor	PyMuPDF, Apache Tika	Text extraction	Handles multiple formats (PDF, DOCX, XML)

6. SCOPE AND LIMITATIONS OF THE STUDY

This paper will discuss the Local RAG deployment in academic libraries in a conceptual and theoretical approach. It is a theoretical paper because it constructs conceptual frames of constructing RAG systems within library settings that examine opportunities of knowledge sharing, improvement in information retrieval and user support services. The investigation has such advantages as increased accessibility to better information and operational efficiency. Since Local RAG is a new technology in library and information science, this underlying concept analysis is the basis of future empirical studies and applications (Khan et al., 2023; Nehra and Bansode, 2024).

The study is a synthesis of existing literature and theoretical knowledge to come up with comprehensive knowledge about Local RAG potential in academic libraries. The methodology will be based on the up-to-date models in related fields that will address the current knowledge appropriately and empirical research on Local RAG in libraries is forming (Bevara et al., 2025). As much as it is an effective approach of identifying the key opportunities and challenges to it bring forth opportunities that can be empirically validated in future through the practical implementation case studies (Mukhopadhyay, 2025). This paper provides a conceptual base that empirical research can be used to enhance knowledge about the Local RAG application in academic libraries (Es et al., 2024).

7. PROSPECTS OF LOCAL RAG IN ACADEMIC LIBRARY SERVICES

Local RAG represents a transformative paradigm for academic library services. As discussed in the literature, Local RAG conceptually addresses long-standing limitations of traditional library systems while aligning with evolving user expectations and institutional governance frameworks.

7.1 Enhanced Accessibility: Local RAG has potential to allows 24/7 chat access to institutional collections without being reliant on the availability of commercial services or operating hour limits. This means that users interact with library resources with their natural language queries at any time and this greatly enhances access to various user groups.

7.2 Information Retrieval: Local RAG systems may offer personalised and context-sensitive service through semantic interpretation and not by simple keyword matching. Systems provide specific suggestions based on real institutional reports that can be accurate and relevant and minimize hallucinations of individual language models.

DOI: 10.63880/jlii.v1i2.49

7.3 Data Sovereignty and Privacy: Local RAG has capabilities to full institutional control of sensitive library data, user queries and collections using local deployment. This could remove privacy risks relating to cloud-based services and be in line with institutional information governance policies as well as compliance requirements.

7.4 Cost-Effectiveness: Local RAG may reduce commercial AI service subscription fee repeated repeatedly, which will lower the costs of long-term operations in comparison to cloud-based solutions. Although it entails significant investment in infrastructure to start up, the lack of recurring vendor charges makes Local RAG financially appealing to resource-limited academic libraries.

7.5 System Customization: Local RAG may be customized to meet the user requirements and workflows based on the needs of individual institutional collections to allow libraries to create solutions that are related to their individual missions and research priorities. Tailoring goes further to support of language specific terminology and integration with the existing library systems.

7.6 Institutional Innovation: The introduction of Local RAG will makes academic libraries innovative institutions which are dedicated to serve the user with the high-level information services. This technology implementation shows that the institution is committed to innovation and it aids in research excellence with the increased access to information.

Table 3. Key Prospects of Local RAG Implementation in Academic Libraries

Prospect Area	Key Benefits	Impact on Library Services	Supporting Evidence
Enhanced Accessibility	24/7 conversational access, natural language queries	Improved service availability for remote and international users	Bevara et al., 2025; Mukhopadhyay, 2025
Information Retrieval	Semantic understanding, personalized recommendations, reduced hallucinations	More accurate and context-aware search results	Gao et al., 2023; Lewis et al., 2020; Bagchi & Mondal 2021
Data Sovereignty and Privacy	Complete institutional control, no external data transmission	Compliance with privacy policies and governance requirements	Xuan, 2025; Xie,2023; Michalak,2024
Cost-Effectiveness	Elimination of recurring subscriptions, long-term savings	Sustainable AI implementation within budget constraints	Xuan, 2025; Rodriguez & Mune, 2022
System Customization:	Tailored to institutional collections and workflows	Services aligned with unique research priorities	Mazumder & Mukhopadhyay, 2024; Oche et al., 2025
Institutional Innovation	Technology-forward positioning, research excellence support	Enhanced institutional reputation and user attraction	Mukhopadhyay, 2025

DOI: 10.63880/jlii.v1i2.49

8. CHALLENGES OF LOCAL RAG IN ACADEMIC LIBRARY SERVICES

Making Local RAG effective in academic libraries could tackle critical issues on technical, organizational and ethical levels. Although these challenges are rather significant they can be overcome with the help of systematic planning and collaboration within the institution.

8.1 Technical Complexity: Local RAG deployment requires technical proficiency in the system integration of machine learning integrations of vectors databases, and continued maintenance. Most of academic libraries do not have technical abilities in house that demand cooperation with the computer science department or external company or technology vendor.

8.2 Computational Resources: Local deployment requires high server resources such as storage capacity of vectors databases in terms of GPU resources and network bandwidth. These infrastructure needs are huge investments of capital and continued operation expenses that might present a challenge to libraries that have small technology budgets.

8.3 Initial Costs: In addition to infrastructure, Local RAG implementation would need development and integration costs such as software licensing, system configuration, data migration and integration with existing library systems. Such initial expenses might prevent libraries implementation even when costs benefits are long-term.

8.4 Staff Training Requirements: The staff at the library need to have training on how to manage the system, troubleshooting, user support and continuous optimization. To answer the questions posed by users and deal with technical problems, the staff should be knowledgeable about RAG technology.

8.5 Quality Assurance: Local RAG systems could be inaccurate or misunderstandings which need human intervention despite being improved over standalone language models. Quality assurances, corrections of errors and improvement processes should be instituted.

8.6 Ethical considerations: Prejudices in training data, privacy of users in using the system and openness in the functioning of the system should be thought of academic libraries have to establish governance systems that deal with ethical use of AI, equity and responsibility.

Table 4. Implementation Challenges and Mitigation Strategies for Local RAG

Challenge Area	Specific Challenges	Mitigation Strategies	Supporting Evidence
Technical Complexity	ML expertise, system integration, maintenance	Collaboration with IT/CS departments, external consultants	Khan et al., 2023; Ilapaka & Ghosh, 2025
Computational Resources	GPU requirements, storage capacity, bandwidth	Phased deployment, infrastructure planning, hybrid solutions	Nehra & Bansode, 2024
Initial Costs	Infrastructure investment, software licensing, integration	Budget planning, pilot projects, grant funding	Comia, 2025; Michalak, 2024

DOI: 10.63880/jlii.v1i2.49

Staff Training	System management, troubleshooting, user support	Structured training programs, documentation, professional development	Xuan, 2025; Oche et al., 2025; Nguyen et al., 2024
Quality Assurance	Potential inaccuracies, misinterpretations	Human oversight, evaluation frameworks, continuous monitoring	Brown et al., 2025; Xu & Loo, 2025
Ethical Considerations	Data bias, privacy protection, transparency	Governance frameworks, ethical AI policies, accountability measures	Wang et al., 2024; Xuan, 2025

9. PPRACTICAL IMPLEMENTATION FRAMEWORK FOR LIBRARY PROFESSIONALS

The implementation of Local RAG needs the methodical planning and gradual implementation in various areas of organizational activity to be successful. this framework incorporates the results of literature analysis, conceptual research and analysis of organizational needs to offer practical guidance to library professionals. Table 5. further details implementation activities across six consecutive phases with key actions and anticipated deliverables being identified to inform resource allocation stakeholder engagement and project management. This is a systematic method that allows libraries to eliminate technical complexity and stay strategically aligned with institutional goals.

Table 5. Six-Phase implementation framework for Local RAG systems in academic libraries, detailing key actions and success indicators for each deployment stage

Implementation Phase	Key Actions	Success Indicators
Phase 1: Planning and Assessment	Conduct feasibility study Evaluate technical infrastructure Assess institutional readiness Analyze budget and resources Identify target users and collections	Feasibility report Needs assessment document Budget allocation plan Stakeholder approval
Phase 2: Technology Architecture	Evaluate open-source LLMs Select vector database solution Choose RAG orchestration Verify system compatibility Design integration protocols	Technology stack documentation Architecture design document Integration roadmap Compatibility assessment report
Phase 3: Pilot Deployment	Design limited-scope pilot Select test collections	Functional pilot system User feedback report

DOI: 10.63880/jlii.v1i2.49

	Implement prototype system Collect performance data Gather user feedback	Performance evaluation Lessons learned document
Phase 4: Capacity Building	Develop training curriculum Conduct staff training sessions Create support documentation Establish troubleshooting protocols Prepare user guides	Trained library personnel Training materials Support documentation User service protocols
Phase 5: Quality and Governance	Implement accuracy testing Establish monitoring systems Develop ethical AI policies Create privacy protocols Design accountability mechanisms	Quality assurance framework Performance monitoring dashboard Governance policy document Compliance protocols
Phase 6: Sustainability and Scaling	Secure ongoing funding Plan system maintenance Establish partnerships Schedule regular updates Enable system expansion	Long-term budget allocation Maintenance schedule Partnership agreements Scaling implementation plan

10. FINDINGS AND DISCUSSION

This study is synthesis of existing literature suggests that Local RAG represents a conceptually viable strategy that may support academic libraries in enhancing information service delivery while maintaining institutional control. This synthesis suggests that Local RAG adoption requires not only technical deployment but also strategic institutional planning, staff capacity building and governance alignment. The findings imply a shift in academic library service models from system-centric search interfaces toward context-aware, conversational and institutionally controlled information environments. Based on the thematic analysis, the study emphasizes the importance of structured and six-phased strategic planning for the adoption of Local RAG in academic library environments. The literature suggests that feasibility assessment plays a key role in understanding institutional technical capacity and infrastructure readiness prior to implementation. Furthermore, staff training programme emerges as a central factor, as trained personnel contribute to effective system management, user support and service optimization.

The reviewed studies also highlight the relevance of governance frameworks that reflect institutional values related to ethical use, privacy protection and accountability. In addition, collaborative engagement with technical experts and peer institutions is viewed as a constructive approach for addressing knowledge gaps. Long-term sustainability of Local RAG initiatives is further supported through planned investment in system maintenance, periodic updates and ongoing professional development, ensuring continued alignment with institutional goals and evolving user expectations.

DOI: 10.63880/jlii.v1i2.49

11. CONCLUSION

Local RAG represents a promising approach for academic libraries seeking to modernize information services while maintaining institutional control, data privacy and independence from commercial artificial intelligence platforms. This study examined the conceptual framework and technical architecture of Local RAG systems and synthesized existing literature to highlight their relevance for academic library environments. Through thematic analysis, the study fulfills its objectives by examining Local RAG architecture, identifying implementation prospects and challenges and proposing a structured six-phase deployment framework for academic libraries.

The study identified key opportunities associated with Local RAG adoption, including improved conversational access to library resources, enhanced relevance of retrieved information and greater flexibility in system customization. The findings also emphasize that effective implementation depends on systematic planning, institutional readiness and alignment with organizational goals. Future research should empirically evaluate Local RAG implementations through pilot deployments in academic libraries, examine user satisfaction and service effectiveness. With appropriate governance and long-term planning, Local RAG has the potential to strengthen the role of academic libraries in supporting teaching, learning and research in the digital era.

REFERENCES

Aboelmaged, M., Bani-Melhem, S., Al-Hawari, M. A., & Ahmad, I. (2024). Conversational AI chatbots in library research: An integrative review and future research agenda. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/09610006231224440

Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health, 4*(6), e0000877. https://doi.org/10.1371/journal.pdig.0000877

Bagchi, A., & Mondal, P. (2021). Understanding information retrieval system in the library of the National Green Tribunal, Eastern Zone Bench, Kolkata: A case study. *Internet Reference Services Quarterly*, 1–13. https://doi.org/10.1080/10875301.2021.1910098

Bevara, R. V. K., Lund, B. D., Mannuru, N. R., Karedla, S. P., Mohammed, Y., Kolapudi, S. T., & Mannuru, A. (2025). Prospects of retrieval augmented generation (RAG) for academic library search and retrieval. *Information Technology and Libraries, 44*(2). https://doi.org/10.5860/ital.v44i2.17361

Brown, A., Roman, M., & Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv*. https://doi.org/10.48550/arxiv.2508.06401

Comia, L. V. (2025). Chatbot for student discipline handbook–related queries: A RAG-based LLM using Llama-3 approach. In *Proceedings of the 11th International Conference on*

DOI: 10.63880/jlii.v1i2.49

Web Research (ICWR) (pp. 306–312). https://doi.org/10.1109/icwr65219.2025.11006165

Cox, A. M., & Mazumdar, S. (2022). Defining artificial intelligence for librarians. *Journal of Librarianship and Information Science, 56*(2). https://doi.org/10.1177/09610006221142029

Cox, A. M., Pinfield, S., & Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. *Library Hi Tech, 37*(3), 418–435. https://doi.org/10.1108/lht-08-2018-0105

Cuconasu, F., Trappolini, G., Siciliano, F., Tonellotto, N., Silvestri, F., Fil-Ice, S., Campagnano, C., & Maarek, Y. (2024). The power of noise: Redefining retrieval for RAG systems. https://doi.org/10.1145/3626772.3657834

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1–12). https://aclanthology.org/2024.eacl-demo.16/

Gao, L., Dai, Z., & Callan, J. (2021). Rethink training of BERT rerankers in multi-stage retrieval pipeline. *arXiv*. https://arxiv.org/abs/2101.08751

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2312.10997

Huang, Y.-H. (2022). Exploring the implementation of artificial intelligence applications among academic libraries in Taiwan. *Library Hi Tech, 42*(3), 885–905. https://doi.org/10.1108/lht-03-2022-0159

Ilapaka, A., & Ghosh, R. (2025). A comprehensive RAG-based LLM for AI-driven mental health chatbot. In *Proceedings of the 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)* (pp. 1–5). https://doi.org/10.1109/ichora65333.2025.11017017

Issifu, I., Abdul, B., & Wumbie, M. (2022). Information retrieval in special libraries and its challenges. *IJIRAS*. https://www.ijiras.com/2022/Vol_9-Issue_1/paper_12.pdf

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12). https://doi.org/10.1145/3571730

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv*. https://doi.org/10.48550/arXiv.2310.06825

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv*. https://arxiv.org/abs/1702.08734

DOI: 10.63880/jlii.v1i2.49

Khan, R., Gupta, N., Sinhababu, A., & Chakravarty, R. (2023). Impact of conversational and generative AI systems on libraries: A use case large language model (LLM). *Science & Technology Libraries, 43*(4), 319–333. https://doi.org/10.1080/0194262x.2023.2254814

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*. https://doi.org/10.48550/arXiv.2005.11401

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv*. https://doi.org/10.48550/arXiv.2305.11747

Lin, J., Nogueira, R., & Yates, A. (2021). Pretrained transformers for text ranking: BERT and beyond. *arXiv*. https://arxiv.org/abs/2010.06467

Mazumder, J., & Mukhopadhyay, P. (2024). Designing question-answer based search system in libraries: Application of open-source retrieval augmented generation (RAG) pipeline. *Journal of Information and Knowledge*, 255–260. https://doi.org/10.17821/srels/2024/v61i5/171583

Michalak, R. (2024). Fostering undergraduate academic research: Rolling out a tech stack with AI-powered tools in a library. *Journal of Library Administration, 64*(3), 335–346. https://doi.org/10.1080/01930826.2024.2316523

Mukhopadhyay, P. (2025). Designing conversational search for libraries: Retrieval augmented generation through open-source large language models. *DESIDOC Journal of Library & Information Technology, 45*(2), 109–115. https://doi.org/10.14429/djlit.20206

Nehra, S. S., & Bansode, S. Y. (2024). Exploring the prospects and perils of integrating artificial intelligence and ChatGPT in academic and research libraries. *Journal of Web Librarianship, 18*(3), 111–132. https://doi.org/10.1080/19322909.2024.2390413

Nguyen, L. T. K., Pham, L. D., & Nguyen, H. N. (2024). uMentor: LLM-powered chatbot for harnessing technology books in digital library. In *Communications in Computer and Information Science* (pp. 232–244). https://doi.org/10.1007/978-3-031-70248-8_18

Nirudi, Y., & Parichi, R. (2025). Artificial intelligence in libraries: An overview. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5080670

Oche, A., Folashade, A., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *arXiv*. https://arxiv.org/pdf/2507.18910v1

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Prajapati, F. (2025). A new concept in the modern library: Steps toward a paperless age. *International Journal of Research in Library Science, 11*, 38–42. https://doi.org/10.26761/ijrls.11.3.2025.1901

DOI: 10.63880/jlii.v1i2.49

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). https://doi.org/10.18653/v1/d19-1410

Rodriguez, S., & Mune, C. (2022). Uncoding library chatbots: Deploying a new virtual reference tool at the San José State University Library. *Reference Services Review, 50*(3/4), 392–405. https://doi.org/10.1108/rsr-05-2022-0020

Roy, P., Kumar, S. K. S., & Satija, M. P. (2012). Problems in searching online databases: A case study of select central university libraries in India. *DESIDOC Journal of Library & Information Technology, 32*(1), 59–63. https://doi.org/10.14429/djlit.32.1.1407

Saha, B., & Saha, U. (2024). Enhancing international graduate student experience through AI-driven support systems: A LLM and RAG-based approach. In *Proceedings of the International Conference on Data Science and Its Applications (ICoDSA)* (pp. 300–304). https://doi.org/10.1109/icodsa62899.2024.10651944

Sihaloho, H. A., Samosir, F. T., & Valentino, R. A. (2025). The implementation of AI-based information retrieval system at the University of North Sumatera Library. *Khizanah Al-Hikmah: Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan, 13*(2), 323–337. https://doi.org/10.24252/v13i2a11

Suryavanshi, K., Thikekar, N., Pawar, R., & Ashtekar, S. (2025). Implementation of RAG-based question-answering application. In *Proceedings of the International Conference on Data Science and Business Systems (ICDSBS)* (pp. 1–6). https://doi.org/10.1109/icdsbs63635.2025.11031968

Wang, L., Wan, Z., Ni, C., Song, Q., Li, Y., Clayton, E., Malin, B., & Yin, Z. (2024). Applications and concerns of ChatGPT and other conversational large language models in health care: Systematic review. *Journal of Medical Internet Research, 26*, e22769. https://doi.org/10.2196/22769

Xie, J. (2023). Research on information retrieval service innovation of university library. *SHS Web of Conferences, 169*, 01088. https://doi.org/10.1051/shsconf/202316901088

Xu, C., & Loo, S. (2025). A review of artificial intelligence applications in libraries in Southeast Asia: Where are we now? *Reference Services Review, 53*(1), 66–91. https://doi.org/10.1108/rsr-06-2024-0027

Xuan, W. (2025). An in-house approach to AI: Developing a custom chatbot for library services. *Internet Reference Services Quarterly, 29*(3), 333–345. https://doi.org/10.1080/10875301.2025.2495917

Yrjo Lappalainen, & Narayanan, N. (2023). Aisha: A custom AI library chatbot using the ChatGPT API. *Journal of Web Librarianship, 17*(3), 37–58. https://doi.org/10.1080/19322909.2023.2221477

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T.-S. (2021). Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv*. https://arxiv.org/abs/2101.00774

DOI: 10.63880/jlii.v1i2.49