Article ID No. 2809 DOI No. https://doi.org/10.5281/zenodo.21214828

Conversational Library Search using RAG and MCP: A Conceptual and Comparative Approach

Jhantu Mazumder
Research Scholar,
jhantumazumder2017@gmail.com

Tirtharaj Dasgupta
Research Scholar
tirtharajlisc25@klyuniv.ac.in

Prof. Parthasarathi Mukhopadhyay
Professor, Dept of Library and Information Science,
University of Kalyani, Kalyani,
psm@klyuniv.ac.in

Jhantu Mazumder
Tirtharaj Dasgupta
Parthasarathi Mukhopadhyay

Abstract:

Purpose *- This paper presents a conceptual and comparative approach of Retrieval Augmented Generation (RAG) and Model Context Protocol (MCP) for integrating and designing conversational library search systems within an open-source architecture. The purpose is to demonstrate how these AI technologies address the persistent challenges in library information retrieval, including hallucinations or fabricated information generation, keyword-based search limitations, lack of contextual responses, fragmented information ecosystems across multiple platforms, low accessibility of high-value resources, real-time information access and limited support for complex users' queries.*

Design/methodology/approach *- The paper discusses the technical knowhow of RAG pipeline integration with LLMs and also the MCP-based agentic system design along with required core components, operational workflows etc.*

Findings *- The paper showcases that LLMs generate contextual responses by augmenting through relevant Library resources, while MCP enables secure and standardized interaction between AI agents and library data sources in real-time and both understand and give responses in Natural language. Both of these mechanisms support and ensure data sovereignty, data privacy, real-time access of Library data along with delivering more precise, contextual, and conversational search experiences.*

Research limitations/implications *- More research needs on evaluating these systems through prototype implementation and user studies and measuring the performance across large library collections, and integration with existing Library Management Systems like ILS, LMS and such other e-databases.*

Originality/value *- This paper is the first to theoretically discuss and demonstrate RAG and MCP for library search systems and give emphasis on the open-source architecture, data governance, advantages and practical implementations for libraries, want to use AI-enabled and NLP (Natural Language Processing) based services.*

Keywords *- Agentic AI, Conversational Search, Large Language Models (LLMs), Library Information Retrieval, Model Context Protocol (MCP), Natural Language Processing (NLP), Retrieval Augmented Generation (RAG).*

1. Introduction:

Libraries are the gateway of knowledge hubs which have been evolving from physical to digital collections. It provides access to printed and electronic resources through institutional repositories, Online Public Access Catalogue (OPAC), discovery platforms etc. In spite of having these advancements, users continue to face several issues in finding the relevant information they need (Aboelmaged *et al.*, 2025)[1]. The traditional keyword-based search systems many times fail to understand the semantic intent or context behind the user queries and leave users frustrated with irrelevant results or make them overwhelmed by information overload (Fitch, 2023)[2].

The application of artificial intelligence has significantly changed the mechanism of information storage, access and retrieval and among all, the most significant developments are Large Language Models (LLMs), which proves remarkable efficiencies in understanding and generating natural language. However, the integration of these powerful technologies into library systems presents both opportunities and unique challenges (Cox *et al.*, 2019)[3]; (Asemi *et al.*, 2021)[4] like hallucinations or real-time information access. As solutions and for enhancing the information retrieval better, Retrieval Augmented Generation (RAG) and Model Context Protocol (MCP) have been introduced.

The integration of conversational AI and agentic systems offers a rapid change. Instead of requiring users to formulate precise Boolean queries or learning advanced search techniques, conversational search systems can interact in natural language, understand context, and retrieve information (Adetayo & Oyeniyi, 2023)[5].

This paper discusses two technologies, Retrieval Augmented Generation (RAG) and Model Context Protocol (MCP) within an open-source architecture. RAG addresses the challenge of LLM hallucinations by augmenting it through authoritative library collections (Lewis *et al.*, 2020[6]; Gao

[1]Aboelmaged, M., Bani-Melhem, S., Ahmad Al-Hawari, M., & Ahmad, I. (2025). Conversational AI chatbots in library research: An integrative review and future research agenda. *Journal of Librarianship and Information Science*, *57*(2), 331–347. https://doi.org/10.1177/09610006231224440

[2]Fitch, K. (2023). Searching for meaning rather than keywords and returning answers rather than links. *The Code4Lib Journal*, (57). https://journal.code4lib.org/articles/17443

[3]Cox, A. M., Pinfield, S., & Rutter, S. (2018). 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

[4]Asemi, A., Ko, A., & Nowkarizi, M. (2020). Intelligent libraries: A review on expert systems, artificial intelligence, and robot. *Library Hi Tech*, *39*(2), 412–434. https://doi.org/10.1108/LHT-02-2020-0038

[5]Adetayo, A. J., & Oyeniyi, W. O. (2023). Revitalizing reference services and fostering information literacy: Google Bard's dynamic role in contemporary libraries. *Library Hi Tech News*. https://doi.org/10.1108/LHTN-08-2023-0137

[6]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

et al., 2023[7]). MCP, introduced by Anthropic in November 2024, provides a standardized protocol for AI agents to connect with external tools and data sources and eliminates the integration fragmentation (Hou *et al.*, 2026[8]; Ray, 2025[9]). These technologies enable the library search systems which are conversational, contextually aware, privacy-preserving, and fully under institutional data control.

2. Theoretical Background and Conceptual Foundations:

Artificial Intelligence (AI) has a subset of technologies designed to simulate human intelligence where Machine Learning (ML) and Deep Learning (DL) are two of them. Machine Learning enables systems to learn from data understanding the pattern without explicit programming and Deep Learning (DL) which is a subset of ML, utilizes neural networks in multiple layers to process complex patterns and enable the foundation for modern Natural Language Processing (NLP) mechanism (Wagner *et al.*, 2022)[10].

But the rise of Large Language Models (LLMs) has changed the total scenario. The LLM models such as Mistral, Mixtral, Llama, and Gemini etc., which are trained on vast corpora and have remarkable capabilities in text generation, question answering, and reasoning. These models are having advanced architectures and mechanisms to address long-range dependencies and contextual nuances (Brown *et al.*, 2020)[11]. The availability of open-source LLMs, including Meta's Llama series and Mistral, has opened access to cutting-edge AI capabilities.

2.1 Libraries in the Digital Age: Persistent Challenges

In spite of applying advanced technologies, libraries face several challenges in information search or retrieval process, such as –

NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[7]Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval Augmented generation for large language models: A survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

[8]Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model context protocol (MCP): Landscape, security threats, and future research directions. *ACM Trans. Softw. Eng. Methodol.* https://doi.org/10.1145/3796519

[9]Ray, P. P. (2025). A survey on model context protocol: Architecture, state-of-the-art, challenges and future directions. *TechRxiv*, *2025*(0418). https://doi.org/10.36227/techrxiv.174495492.22752319/v1

[10]Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

[11]Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

(i) **Keyword-Based Search Limitations:** Traditional Integrated Library Management System, Online Public Access Catalogue (OPAC), Institutional Repositories, Discovery systems etc rely on exact keyword matching, which often fails to understand the context of the queries. The users who are lacking domain expertise or the advanced search strategies face challenges to receive relevant results (Li & Coates, 2025)[12].

(ii) **Federated Search System:** Library resources are distributed across multiple platforms which hinder the obtaining of contextual information (Bevara *et al*., 2025)[13].

(iii) **24*7 Reference Services:** Library staff provide responses but cannot offer 24/7 personalized assistance. The chatbots have worked as partial solutions but often lack context and accuracy (Rodriguez & Mune, 2022[14]; Lappalainen & Narayanan, 2023[15]).

(iv) **Hallucinations and Relevancy Concerns:** When LLMs are deployed without grounding with relevant resources, they may generate irrelevant and non-contextual responses which is known as hallucination and that create significant risks in academic and research contexts (Ji *et al*., 2023[16]; Li, 2023[17]).

2.2 Importance of Open-Source and Sovereign Solutions:

The proprietary or closed-source AI services work through external servers which creates concerns about data privacy, data governance, and institutional control. For libraries committed to user privacy, such mechanisms are problematic (Yoon *et al*., 2022)[18]. The open-source architectures solve this gap and offer an alternative. Libraries can deploy these open-source AI technologies on their own infrastructure to maintain complete control over their data.

[12]Li, L., & Coates, K. (2024). Academic library online chat services under the impact of artificial intelligence. *Information Discovery and Delivery*, *53*(2), 192–205. https://doi.org/10.1108/IDD-11-2023-0143

[13]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

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

[15]Lappalainen, Y., & 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

[16]Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, *55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

[17]Li, Z. (2023). The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination. *Nature Machine Intelligence*, *5*(6), 559–560. https://doi.org/10.1038/s42256-023-00672-y

[18]Yoon, J., Andrews, J. E., & Ward, H. L. (2021). Perceptions on adopting artificial intelligence and related technologies in libraries: Public and academic librarians in North America. *Library Hi Tech*, *40*(6), 1893–1915. https://doi.org/10.1108/LHT-07-2021-0229

3. Retrieval Augmented Generation (RAG) in Library Search:

Retrieval Augmented Generation (RAG) represents a rapid change in how LLMs generate responses. Instead of relying solely on the knowledge embedded during ingesting, RAG systems index and retrieve most relevant documents from the knowledge base, rank them as per relevancy and incorporate them into the response generation process (Lewis *et al.*, 2020)[19]. This approach addresses the most critical limitation of standalone LLMs, hallucinations (Gao *et al.*, 2023)[20]; (Chen *et al.*, 2024[21]).

3.1 RAG Architecture for Libraries:

Designing the RAG based architecture for the library search system requires several software components which are given in Table 1.

Table 1. Core Components of an Open-Source RAG based System for Libraries

Component	Role	Tools
RAG Pipeline	Retrieval and response generation processes	LangChain
Vector Database	Stores document embeddings for efficient semantic retrieval	ChromaDB
Embedding Model	Converts textual data into vector representations	intfloat/e5-large-v2
Large Language Model (LLM)	Generates contextual responses from retrieved information	Llama 3, Mistral (via Groq, Ollama)
End User Interface	Provides user interface for query input and interaction	Streamlit
Data Ingestion Tools	Collects, processes, and converts library data into effective format	wget, WARC tools

3.2 Operational Workflow:

The operational workflow of a RAG-based library search system can be conceptualized as a two-stage process-

i) Data Ingestion:

In this stage, relevant library content including website pages, policy documents, database descriptions, and details about the library resources and services are systematically collected and preprocessed. The collected data is then converted into WARC format to ensure efficient

[19] Ibid; See Footnote- 6.
[20] Ibid; See Footnote- 7;
[21] Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in Retrieval Augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

storage, scalability, and structured processing (Mazumder & Mukhopadhyay, 2024)[22]. Therefore, the textual content is divided into small chunks and converted into vector representations using transformer-based embedding models, and stored within a vector database called ChromaDB. Finally, a semantic index is constructed to enable efficient similarity-based retrieval required during user query processing.

ii) Query Processing:

At this stage, the user submits a query in natural language through the end-user interface. The query is converted into a vector representation using the same embedding model. The vector database then performs a similarity search to retrieve the top-k most relevant documents based on semantic proximity and send them to LLMs. These retrieved documents serve as contextual grounding for the response generation process. The Large Language Model (LLM) then generates a contextual and personalised response by incorporating the retrieved information.

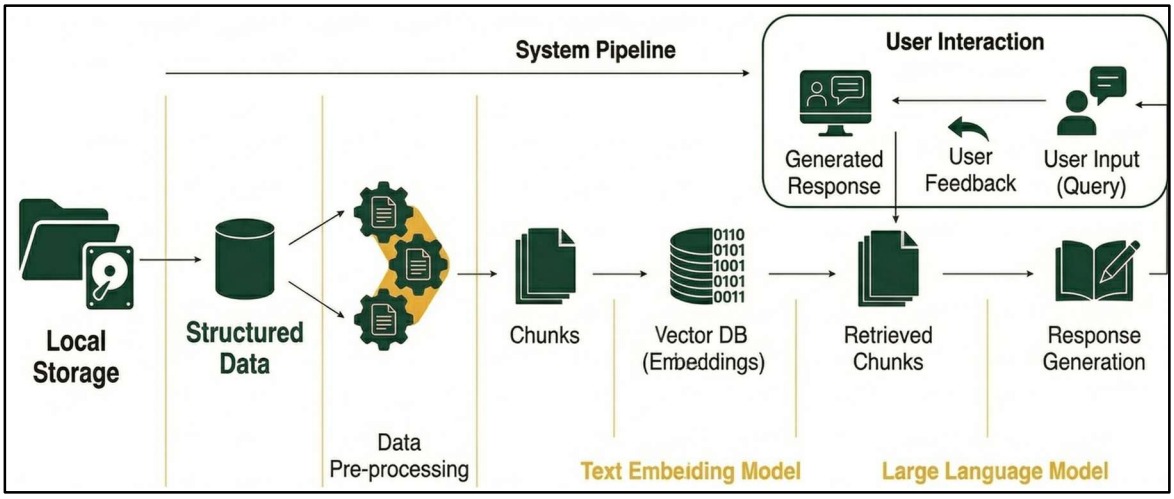

Figure 1. Semantic search through RAG

3.3 Practical Implementation Examples:

The experiments and practical applications of RAG-based library search systems have been demonstrated in recent research. Mazumder and Mukhopadhyay (2024)[23] designed a prototype using LangChain, ChromaDB, and Llama3-70b that successfully answered queries about the Chandrayaan-3 mission. The study found that standalone LLMs produced confidently incorrect answers but when integrated with RAG, the system consistently provided accurate and

[22] 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

[23] Ibid; See Footnote- 22;

contextualized responses. Similarly, Yang and Zhang (2024)[24] experimented and demonstrated RAG integration with an institutional repository (used DSpace-7) and enabled conversational search for institutional collections. Bevara *et al.* (2025)[25] explored the considerable potential of RAG in academic library search and highlighted its efficiency for semantic indexing and real-time query processing.

4. Model Context Protocol (MCP) in Library Search:

The Model Context Protocol (MCP) was introduced by Anthropic in November 2024 that addresses a critical challenge in agentic AI by integrating AI agents and external tools or data sources (Hou *et al.*, 2026[26]; Ehtesham *et al.*, 2025[27]). Before MCP, each tool required custom implementation resulting in an $m \times n$ integration problem where m agents needed separate connections to n resources. MCP provides a standardized and open protocol that enables AI agents to connect seamlessly with –

i) **Tools:** Executable functions that extend agent capabilities (database queries, API calls);
ii) **Data Sources:** Files, databases, and repositories containing authoritative information;
iii) **Workflows:** Structured prompts and processes that guide agent behavior.

The protocol works through a three-entity architecture-

i) **MCP Host:** The application (chat interface, code editor) providing user interaction;
ii) **MCP Client**: Code within the host that manages server connections;
iii) **MCP Server:** Implements external functionalities and exposes them to clients.

4.1 MCP Architecture for Libraries

The architecture of MCP based systems in libraries is based on a standardized client-server protocol that facilitates agent-driven interaction with tools, data sources, and library systems, as summarized in Table 2.

Table 2. MCP Architecture Components for Library Implementation

Component	Description	Library Application
MCP Host	User-facing application interface	LibreChat, Claude Desktop, custom frontends
MCP Client	Integration layer that connects host to MCP servers	Embedded within the host application

[24]Yang, L., & Zhang, Z. (2024). Using RAG, LLMs, and LangChain to create a query application on open repositories: Demo application on DSpace-7. *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 1–2. https://doi.org/10.1145/3677389.3702488

[25] Ibid; See Footnote- 13

[26] Ibid; See Footnote- 8;

[27]Ehtesham, A., Singh, A., Gupta, G. K., & Kumar, S. (2025). *A survey of agent interoperability protocols: Model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP)* (arXiv:2505.02279). arXiv. https://doi.org/10.48550/arXiv.2505.02279

MCP Server	Executes library-specific tools and functionalities	VuFind MCP Server, DSpace MCP Server, ILS connectors
Data Layer	Defines client–server communication using JSON-RPC	Standardized message exchange format
Transport Layer	Handles communication channels (e.g., stdio, SSE)	Supports local and remote system connections
Authentication	Manages secure access using OAuth 2.x protocols	Ensures authorized access to library systems and resources

4.2 Operational Workflow:

In an MCP-enabled library search system, the process begins when a user submits a natural language query through the MCP host interface, such as LibreChat. The Large Language Model (LLM) then interprets the query and determines the appropriate actions required to fulfill the user's request. Based on this analysis, the MCP client communicates with the relevant MCP servers, which are responsible for executing specific tool functions, such as retrieving records via REST API-based services. The results generated by the MCP server are then returned to the LLM, which synthesizes the information into a contextual and relevant response. Finally, the user receives a comprehensive and personalized response with real-time information access from the library systems.

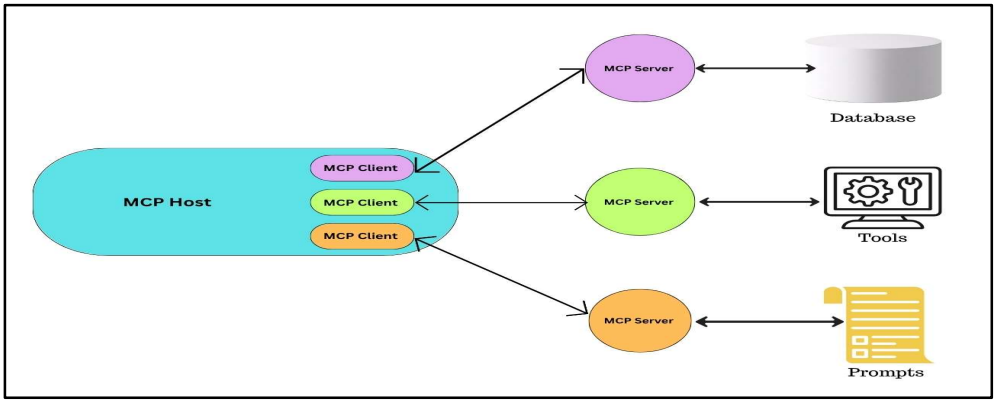

Figure 3. Semantic Search through Model Context Protocol (MCP)

4.3 Practical Implementation Examples:

Recent studies are identified which demonstrate the practical implementation of Model Context Protocol (MCP) in library systems and highlight its potential, those are –

(i) **Conversational Discovery:** Neogi *et al.* (2025)[28] implemented an MCP-based conversational retrieval system for VuFind which understands natural language queries, retrieves records

[28]Neogi, M., Dasgupta, T., & Mukhopadhyay, P. (2025). NLP-based library retrieval: Integrating VuFind with LLM through MCP middleware. *Indian Journal of Information Library & Society*, *38*(1–2), 1–19.

by author, title, and subject and generates responses in natural language. The system successfully demonstrated LLM capability to analyze queries using REST APIs, and return relevant results.

(ii) **Institutional Repository Integration:** Dasgupta *et al.* (2025)[29] developed an MCP server for DSpace which enables conversational retrieval from institutional repositories. The system successfully returned records using natural language search.

(iii) **Open-Source Framework:** Dasgupta and Mukhopadhyay (2025)[30] implemented a fully open-source MCP-based search system using MCP CLI and open-source LLMs via Ollama, overcoming limitations of closed-source dependencies.

(iv) **Special Collections Access:** Jana and Rout (2026)[31] demonstrated MCP application in retrieving records from specialized collections (Sambalpuri Sarees) with domain-specific metadata.

(v) **Next-Generation Library Service Platforms:** Liu *et al.* (2025)[32] conceptualized A-LSP, a three-tier agentic architecture for library service platforms built on MCP principles.

5. Comparative Approach: RAG and MCP in Library Contexts:

In the context of library information retrieval, it is essential to comparatively examine RAG and MCP in terms of their architecture, data handling, operational workflows, and applicability to better understand their functional roles.

Table 3. Comparative Approach of RAG and MCP based Model

Aspect	RAG Model (Retrieval Augmented Generation)	MCP Model (Model Context Protocol)
Core Function	Knowledge grounding through semantic retrieval	Standardized agent–tool integration
Data Source	Static or curated documents (e.g., policies, metadata, descriptions)	Live systems (e.g., catalogs, ILS, IRs, databases)
Response Basis	Pre-ingested and indexed content	Real-time data via API calls
Implementation Complexity	Moderate (pipeline setup, embedding, indexing)	Moderate to high (server development and integration)
Currency of	Depends on ingestion frequency	Real-time and continuously updated

[29]Dasgupta, T., Neogi, M., & Mukhopadhyay, P. (2025). Enhancing DSpace with large language models: Designing an integration framework using the model context protocol. *Proceedings of KEDLD-2025*, 235–245.

[30] Dasgupta, T., & Mukhopadhyay, P. (2025). Searching the library through commands: Implementing an open source command line-based conversational library search system using model context protocol. *Proceedings of LAB International Conference*, 249–261.

[31]Jana, A., & Rout, R. (2026). Sambalpuri sarees in the digital loom: Metadata schema and natural language query interface design. *Journal of Library Metadata*, *0*(0), 1–20. https://doi.org/10.1080/19386389.2026.2636260

[32]Liu, W., Zhang, L., Ji, T., & Chen, X. (2025). Shaping the smart libraries with AI AI: An agent-based based, next-generation library service platform. *Journal of Library & Information Science in Agriculture*, *37*(5), 15. https://doi.org/10.13998/j.cnki.issn1002-1248.25-0379

Information		
Action Capability	Limited to information retrieval	Supports transactional operations (e.g., holds, renewals)
Security Model	Data remains within institutional infrastructure	OAuth-based authentication and controlled access
Cost Structure	Storage and inference costs	Storage, inference, and server hosting costs
Best Suited For	Policy Q&A, reference services, and user guidance	Catalog search, circulation tasks, and live data access

6. Future Research Directions:

Though RAG and MCP demonstrate significant potential in designing conversational library search, further research is required to address key technical, operational, and integration challenges. While RAG requires research on the areas of performance benchmarking, multi-lingual capabilities, dynamic knowledge updates and evaluation metrics, MCP requires research on its security vulnerabilities, performance optimization, integration of data-sources and integration of other agentic standards in libraries.

7. Conclusion:

This study provides a conceptual and comparative analysis of Retrieval Augmented Generation (RAG) and Model Context Protocol (MCP) within the context of conversational library search systems. The findings indicate that both models address the issues of information retrieval in Library search. RAG enhances the capability by augmenting authoritative and curated knowledge base and reduces hallucinations. Similarly, MCP enhances the capability of search systems by enabling real-time interaction with external tools or data sources.

Accordingly, this paper considers RAG and MCP not as competing or uniformly integrated systems, but as cutting-edge technologies for enhancing the library search systems. This perspective changes the understanding of library information retrieval from static, keyword-based processes to contextual conversational systems. The conceptual contribution of this study lies in clarifying the distinct functional roles of augmentation-based and agent-based mechanisms and highlighting their potentials when applied selectively.

References:

Aboelmaged, M., Bani-Melhem, S., Ahmad Al-Hawari, M., & Ahmad, I. (2025). Conversational AI chatbots in library research: An integrative review and future research agenda. *Journal of Librarianship and Information Science*, *57*(2), 331–347. https://doi.org/10.1177/09610006231224440

Adetayo, A. J., & Oyeniyi, W. O. (2023). Revitalizing reference services and fostering information literacy: Google Bard's dynamic role in contemporary libraries. *Library Hi Tech News*. https://doi.org/10.1108/LHTN-08-2023-0137

Asemi, A., Ko, A., & Nowkarizi, M. (2020). Intelligent libraries: A review on expert systems, artificial intelligence, and robot. *Library Hi Tech*, *39*(2), 412–434. https://doi.org/10.1108/LHT-02-2020-0038

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, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in Retrieval Augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Cox, A. M., Pinfield, S., & Rutter, S. (2018). 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

Dasgupta, T., & Mukhopadhyay, P. (2025). Searching the library through commands: Implementing an open source command line-based conversational library search system using model context protocol. *Proceedings of LAB International Conference*, 249–261.

Dasgupta, T., Neogi, M., & Mukhopadhyay, P. (2025). Enhancing DSpace with large language models: Designing an integration framework using the model context protocol. *Proceedings of KEDLD-2025*, 235–245.

Ehtesham, A., Singh, A., Gupta, G. K., & Kumar, S. (2025). *A survey of agent interoperability protocols: Model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP)* (arXiv:2505.02279). arXiv. https://doi.org/10.48550/arXiv.2505.02279

Fitch, K. (2023). Searching for meaning rather than keywords and returning answers rather than links. *The Code4Lib Journal*, (57). https://journal.code4lib.org/articles/17443

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval Augmented generation for large language models: A survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model context protocol (MCP): Landscape, security threats, and future research directions. *ACM Trans. Softw. Eng. Methodol.* https://doi.org/10.1145/3796519

Jana, A., & Rout, R. (2026). Sambalpuri sarees in the digital loom: Metadata schema and natural language query interface design. *Journal of Library Metadata*, *0*(0), 1–20. https://doi.org/10.1080/19386389.2026.2636260

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, *55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

Lappalainen, Y., & 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

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. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, L., & Coates, K. (2024). Academic library online chat services under the impact of artificial intelligence. *Information Discovery and Delivery*, *53*(2), 192–205. https://doi.org/10.1108/IDD-11-2023-0143

Li, M., Kilicoglu, H., Xu, H., & Zhang, R. (2025). BiomedRAG: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, *162*, 104769. https://doi.org/10.1016/j.jbi.2024.104769

Li, Z. (2023). The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination. *Nature Machine Intelligence*, *5*(6), 559–560. https://doi.org/10.1038/s42256-023-00672-y

Liu, W., Zhang, L., Ji, T., & Chen, X. (2025). Shaping the smart libraries with AI AI: An agent-based based, next-generation library service platform. *Journal of Library & Information Science in Agriculture*, *37*(5), 15. https://doi.org/10.13998/j.cnki.issn1002-1248.25-0379

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

Neogi, M., Dasgupta, T., & Mukhopadhyay, P. (2025). NLP-based library retrieval: Integrating VuFind with LLM through MCP middleware. *Indian Journal of Information Library & Society*, *38*(1–2), 1–19.

Radosevich, B., & Halloran, J. (2025). *MCP safety audit: LLMs with the model context protocol allow major security exploits* (arXiv:2504.03767). arXiv. https://doi.org/10.48550/arXiv.2504.03767

Ray, P. P. (2025). A survey on model context protocol: Architecture, state-of-the-art, challenges and future directions. *TechRxiv*, *2025*(0418). https://doi.org/10.36227/techrxiv.174495492.22752319/v1

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

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Yang, L., & Zhang, Z. (2024). Using RAG, LLMs, and LangChain to create a query application on open repositories: Demo application on DSpace-7. *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 1–2. https://doi.org/10.1145/3677389.3702488

Yoon, J., Andrews, J. E., & Ward, H. L. (2021). Perceptions on adopting artificial intelligence and related technologies in libraries: Public and academic librarians in North America. *Library Hi Tech*, *40*(6), 1893–1915. https://doi.org/10.1108/LHT-07-2021-0229

Zheng, Q., Chen, M., Park, H., Xu, Z., & Huang, Y. (2025). Evaluating non-AI experts' interaction with AI: A case study in library context. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, 1–20. https://doi.org/10.1145/3706598.3714219