Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39322

Applications of Large Language Model Agents in Scholarly Information Retrieval at University Libraries

Beihang Guo

Department of Library and Information Science, Universiti Malaya, Kuala Lumpur, Malaysia
gbeihang@gmail.com

Abstract. Large language model (LLM) agents can translate the natural language needs of researchers into a set of tasks: searching, screening, citation expansion, and evidence integration. The paper explores their use for academic information retrieval services in university libraries from three perspectives: technical mechanisms, empirical evidence, and deployment constraints. The paper compares retrieval-augmented generation (RAG), agentic retrieval with tools, and agentic retrieval workflows, discussing datasets, metrics, retrieval scope, and supervision. Previous work has shown that increased workflow complexity does not necessarily improve precision, evidence reliability, or coverage, although iterative querying and citation expansion can improve coverage in selected tasks; additional processing may also increase operating costs. Open benchmark data and datasets based on arXiv cannot be directly ported to library environments with licensed databases, heterogeneous collections, copyright restrictions, and librarian review. Thus, the paper proposes: task-specific automation, explicit stopping conditions, provenance recording, source verification, and escalation to human staff. Future evaluation should include retrieval metrics, besides accessibility, service resolution, review burden, privacy, copyright compliance and user satisfaction. University libraries can employ the following actions to leverage LLM agents as assistive tools that support professionals rather than replace them.

Keywords: Large language model agents, University libraries, Scholarly information retrieval, Tool use, Retrieval-augmented generation

1. Introduction

In university libraries, scholarly information retrieval is more than a keyword match. It also involves the choice of database, bibliographic verification, access to full text, tracking of references and interpretation of evidence. The traditional keyword searching and Boolean retrieval are good tools for user control but they do require the user to know the discipline terminology, the search fields and the database rules. Large language models (LLMs) can understand natural-language requests, yet may also generate references that are unavailable, or make claims that are not supported by the cited sources. Retrieval-augmented generation (RAG) reduces dependence on parametric knowledge with external documents. However, retrieving relevant passages does not guarantee full search and accurate comprehension of evidence [1, 2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39322

LLM agents that combine problem understanding, tool usage, and sequential decision making enhance this capability. Yao et al. proposed Reasoning and Acting (ReAct) [1] which combines reasoning trace with actions and observations to enable multistep tasks. Wang et al. provided an overview of LLM-based autonomous agents from the perspectives of architecture, construction methods, applications, and evaluation, including an outline of the key design considerations and research challenges relevant to search agents [3]. In library field, Zhu designed a multi-agent system for subject service, which mainly realized the functions of indicator automatic extraction and report generation [4]. Zhang proposed a framework and implementation process for the artificial intelligence (AI) virtual librarian services in university libraries, which integrate the question-answering repository, the document repository and the application repository, but still can access human librarians [5]. Wang studied the embedded knowledge services using LLMs, and separated knowledge retrieval, question-answering and knowledge discovery into three independent service tasks [6]. All of these studies suggest that agents can be useful for assisting with discovery, reference and knowledge organization, but performance on algorithmic benchmarks is not synonymous with effectiveness in real library services.

Capabilities of agents can also be defined by studies on retrieving scholarly papers. PaSa, an LLM-based paper-search agent, performs multi-step operations such as query generation, paper search, citation expansion, and relevance screening [7]; Scholar Paper Retrieval (SPAR) performs query understanding, retrieval, relevance screening, query evolution, and reranking [8]. However, these experiments are largely limited to open resources, selected disciplines or constructed queries. They are very different from the variety of databases, Chinese language materials and diverse collections found in university libraries. Thus, in translation into operational services important factors like coverage of resources, permissions of the interface, copyright issues, evaluation criteria and human staff must be taken into account.

The paper is a narrative review of the Chinese- and English-language literature from the library and information science journals. It contains representative publications on agents, retrieval technologies or university library services. Studies that clearly report tasks, data, metrics or service cases are given particular attention. Works on general text generation (without contextual/retrieval data) are not included. It examines first retrieval control mechanisms, then existing evidence for RAG, tool use and multistep agentic retrieval, and finally service scenarios, evaluation techniques and deployment limitations for university libraries.

2. Concepts and analytical framework

Large language models are responsible for language understanding and generation, while agents provide tools, state, and action rules on top of the LLM and choose next action based on the LLM output. Planning, memory and reflection can reinforce a system, but their presence alone is insufficient to define an agent. The real issue is whether or not the system can choose tools, adapt queries and deal with exceptions as the task changes. A fixed workflow should not be automatically assumed to be a very autonomous agent if there is a language model at the end of it [3]. There are different levels of automation for known-item searching, topic discovery, and synthesis of viewpoints, all of which span from bibliographic records and abstracts to full text and citation relationships within the university library [6].

This paper describes scholarly retrieval as a process of ongoing retrieval, acquisition, screening and interpretation of academic evidence. The production of an answer is not enough to complete the task. Evaluation should also take into account whether candidate publications are relevant, if key publications are missing and if the conclusions are supported by the evidence. RAG first fetches

DOI: 10.54254/2755-2721/2026.39322

external evidence and then structures the answer via a generative model [9]. Other tasks in its implementation include document parsing, segmentation, indexing, reranking and context construction. Comparative evidence shows that the performance of RAG is influenced by the interaction between these components and not just the generative component itself [10]. Despite these efforts, one can not always depend on RAG to provide accurate citations or to retrieve all the relevant material, as there may be omissions, segmentation errors, and contextual noise in the final response.

The use of tools turns a task into executable operations like selecting a database, specifying a query, specifying a date range, and filtering documents based on the document type. API-Bank experiments demonstrate that the accuracy of sequential tool use is dependent on model type and the type of task [11]. Based on this, the paper categorizes the various retrieval approaches into conventional retrieval, predefined RAG pipelines, tool-using agents, and agentic RAG, depending on the level of control over retrieval. The comparison focuses on query control, follow-up queries, stopping conditions, and whether errors can be traced in an operational record.

3. Technical approaches and evidence

Traditional retrieval needs the users to constantly modify the search statement and filters. In agentic retrieval, in contrast, the model decides what to do next based on the information it has already acquired. InteR is an iterative retrieval framework rather than a fully autonomous agent. It leverages retrieved content to improve generation and generated knowledge to facilitate retrieval [12]. ReAct takes actions on observations and PaSa takes actions on paper searching, citation expansion and stopping decisions [1, 7]. Known-item retrieval is easier if one knows, for example, title or author. Agents improve the ability to evolve queries and, in the case of complex topics, to supplement the search.

Standard predefined RAG pipelines usually follow a specific order of retrieval, reranking and generation. They can be used as a way to structure local knowledge bases, the resources related to them and questions about them. Different RAG configurations do not yield consistent results on general retrieval datasets. In Table 1, Wang et al. reported higher normalized discounted cumulative gain at rank 10 (nDCG@10) for Hypothetical Document Embeddings (HyDE) than for the base retriever on the Text REtrieval Conference (TREC) Deep Learning 2019 (DL19) and 2020 (DL20) passage-ranking datasets, whereas query rewriting and query decomposition did not improve the scores [13]. These are statistics that measure the rank of a retrieved result, not the accuracy of results.

Table 1. Ranking performance of different retrieval configurations

Retrieval configuration	DL19 nDCG@10	DL20 nDCG@10
LLM-Embedder	0.7020	0.6876
With query rewriting	0.6789	0.6562
With query decomposition	0.6610	0.6495
With HyDE	0.7544	0.7394
With hybrid retrieval	0.7250	0.6980

While HyDE was better on ArguAna and FiQA, InteR actually performed better on DL19 and DL20, but not in all datasets, suggesting that its improvement was not consistent across different

DOI: 10.54254/2755-2721/2026.39322

datasets [12]. ReAct showed the potential of combining reasoning and action for HotpotQA and FEVER, but it failed to beat all of the baselines [1]. Technical approaches cannot be compared on a single score; results should be compared to the data source, the task type, and the comparison baseline.

PaSa is composed of two parts: Crawler and Selector. Its AutoScholarQuery training, validation, and test sets include 33,551, 1,000 and 1,000 queries, respectively, and RealScholarQuery includes 50 queries. The experiments are mainly on the field of artificial intelligence and papers in arXiv [7]. As seen in Table 2, PaSa-7b outperforms the Google plus GPT-4o combination in terms of Recall@20 and Recall@50 on both test set and real-query set. This finding indicates that query expansion and multiple queries will be helpful for enhancing document coverage for the selected tasks. Comparisons are, however, required to differentiate between final recall, Crawler recall and Recall@k.

Table 2. Scholarly retrieval results reported in the PaSa study

Dataset	Method	Recall@20	Recall@50
AutoScholarQuery test set	Google + GPT-4o	0.1921	0.2450
AutoScholarQuery test set	PaSa-7b	0.5301	0.6334
RealScholarQuery	Google + GPT-4o	0.2020	0.2573
RealScholarQuery	PaSa-7b	0.5798	0.6563

SPAR fuses the concepts of query understanding, retrieval, relevance assessment, query evolution and reranking. There are 50 queries in the SPARBench: 35 in computer science and 15 in biomedicine. The queries were not extracted from the original user logs [8] and 560 papers were identified by human review. Table 3 demonstrates that both the precision and the recall of SPAR are better than the two comparison methods on this benchmark, with the F1 score being the highest. However, the observed differences may also reflect differences in data sources and evaluation stages, not only system architecture. The PaSa recall reported in the SPAR paper refers to Crawler recall, not PaSa's final recall [7, 8].

Table 3. SPAR results on SPARBench

Method	Precision	Recall	F1
Google search restricted to arXiv	0.2180	0.2800	0.2451
PaSa	0.1076	0.1009	0.1041
SPAR	0.2932	0.3103	0.3015

Agentic RAG can analyze the available evidence and determine if the question needs to be broken down into a series of questions, if further evidence should be retrieved, or if the answer should be modified, adding action traces and feedback to evaluation [14]. As shown in Table 4, suitable tasks and key checks for four approaches are shown as a function of their control mechanisms [1, 7, 10, 14]. Simple questions can be answered by retrieving information directly from a reliable source whereas complex topics might warrant multi-step searching. Further processing does not yield additional evidence once the available evidence is what it is, and further processing will incur additional costs and risk of accumulating errors.

DOI: 10.54254/2755-2721/2026.39322

Table 4. Suitable conditions for different retrieval approaches

Approach	Retrieval control	Suitable task	Key checks
Conventional retrieval	User-defined queries and filters	Known-item and precise field searching	Query coverage
Predefined RAG pipeline	Generation after fixed-pipeline retrieval	Local knowledge-base question answering	Recall quality and evidence alignment
Tool-using agent	Tool and action selection based on feedback	Sequential cross-source retrieval	Tool-selection and parameter errors
Agentic RAG	Dynamic supplementary search and evidence assessment	Complex topic retrieval and synthesis	Stopping conditions and error accumulation

ReAct, InteR, PaSa and SPAR are evaluated on question-answering, general retrieval, or scholarly paper collections, many of which rely heavily on arXiv. University libraries must also address licensed-database permissions, heterogeneous collections, copyright restrictions, and librarian review. Technical potential is therefore assessed through public benchmarks, but this does not necessarily translate to actual accurate, accessible and cost-effective services.

4. Applications and constraints in university libraries

4.1. Service scenarios and human oversight boundaries

The library research is related to the service organization and integration in the library with operational processes. The direct answer-retrieval augmented-answer approach has been used in the retrieval-augmented thinking literature [15] and in knowledge service systems of libraries using LLM technology [16]. The use of repositories of documents and applications, as well as channels for human intervention with the ability to answer questions, is the basis for the application of AI virtual librarians in university libraries [5]. Another dimension of intelligent reference services has been the study of the interconnections among knowledge bases, business systems, and service workflows [17, 18]. Then, the studies on service paradigms and application scenarios address the issues of technical effectiveness, human values, multilingual services and ecosystem coordination [19, 20]. From these practices one can note that the library agent must do more than just provide answers, it must integrate provenance, access policies, exception handling and human verification into the service chain.

There are studies of embedded knowledge services that consider retrieval of resources, question answering and knowledge discovery as independent tasks [6]. However, the extraction of indicators and the generation of reports can be automated by means of multi-agent subject services, and data from some specialized databases still need to be uploaded manually [4]. The associated experiment showed that the multi-agent system saved substantial time, but slightly less data accuracy and depth of argument than the human service. The comparison is shown in Table 5. Accuracy was defined as the percentage of correctly reported indicators and argument depth was evaluated using blinded expert ratings on a five-point scale [4]. The result suggests that the workflow is feasible, but the study did not report the sample size, repeated runs, or maintenance costs.

DOI: 10.54254/2755-2721/2026.39322

Table 5. Comparison of a subject-analysis agent and human service

Service mode	Data accuracy	Argument depth	Time
Multi-agent	89.3%	4.3	8 min
Human service	94.7%	4.6	180 min

Some useful capability references for libraries could be found in both PaSa and SPAR, but real queries of PaSa focus on artificial intelligence and real queries of SPAR focus on computer science and biomedicine [7, 8]. How to adapt to the terminology and resources of Chinese theses, humanities and social sciences, and local collections remains to be seen. Agents can be used to extract the research objects, methods, data and conclusions from the source text, and also find out where the object, method, data and conclusions can be found. In the first stage of relevance screening, however, quality judgments about the research cannot be substituted and having access to an abstract does not automatically allow the full text to be acquired and processed.

4.2. Evaluation and deployment constraints

The accuracy of question answering is not sufficient to be used as the basis of the evaluation of service tasks. Paper discovery is about coverage and omissions, screening and synthesis about field extraction and evidence alignment, reference services about problem resolution and passing complex tasks to staff, resource navigation about successful completion, and subject services about indicator accuracy and analytical quality. Table 6 [4, 5, 7, 8] summarizes the main application, evaluation priorities and the human responsibilities. First, libraries can automate activities with clear rules and stable sources of information and then expand their use of agents based on the process records.

Table 6. Application tasks and evaluation priorities

Application	Output	Evaluation priority	Primary human responsibility
Paper discovery	Candidate papers and search records	Recall and relevance	Check coverage and omissions
Screening and synthesis	Field extraction and evidence alignment	Extraction accuracy and source matching	Decide inclusion and interpret evidence
Reference service	Answers and resource links	Accuracy and problem-resolution rate	Handle complex inquiries
Resource navigation	Database selection and operating path	Task completion rate	Handle permissions and exceptions
Subject services	Monitoring results and analytical reports	Indicator accuracy and analytical quality	Apply disciplinary judgment

Scholarly retrieval hallucinations are spurious references and unsupported claims without real sources. Even in case of external retrieval, errors can still happen with people, works or citations. Therefore, the relation between content and source has to be verified [15]. Multistep workflows can also compound misconceptions about the question. The system should therefore retain the original

DOI: 10.54254/2755-2721/2026.39322

request, query conditions, result lists, location of evidence and exception records. The multiple agent agreement is not a substitute for the independent verification of the source.

Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of all relevant documents that are retrieved. The F1 score is the harmonic mean of precision and recall. While the increased number of citations may improve coverage, it can also introduce noise and boost the costs of using the tool [7, 8]. Additionally, there are few open application programming interfaces (APIs) and APIs that exist in many licensed databases may have restricted authorization and copyright policies. If the system cannot obtain a resource directly from it, it should support searching the resource by a librarian or importing it from a file and report on any missing information. The topics studied, the queries and reading preferences can also provide indications of unpublished work and therefore the scope of data processing and data storage should be defined in advance of using external models.

Agentic RAG needs to be evaluated with more than recall scores, and evaluation should include number of tool calls, runtime, maintenance costs, and librarian review time. The choice of technology should therefore fulfill four conditions: availability of resources, traceability of processes, fulfillment of copyright requirements and clear responsibility.

5. Directions for future research

Previous studies have demonstrated the feasibility of agentic retrieval, virtual librarians and subject services. However, these approaches have not been fully tested in a real library environment. Librarians and researchers could then work together to assemble the task sets, which are anonymized, and compare the performance of fixed pipeline RAG against agentic RAG on the same model, data, and budget for tool uses. Studies of these should examine whether the use of cross-database retrieval, additional search, stop and escalation to human staff actually improves the results obtained [7, 14].

Alternatively, the capabilities of single- and multi-agent systems for the same tasks can be evaluated to see whether sharing retrieval, field extraction and source verification tasks reduces the burden on the librarian and to establish guidelines for reconciling conflicting agent conclusions [4]. A third approach is to look at the licensed, non-open databases: bibliographic completeness, copyright considerations, and service timeliness of file import, librarian-assisted access, and programmable interfaces [5].

Relevance, recall, cost and other factors should be extended to include source accessibility, effectiveness of human escalation, copyright compliance, user satisfaction, efficiency of material acquisition, repetitive librarian work and burden of correcting errors. The only way to determine if the benefits of agents outweigh their added maintenance and governance costs is to assess them against real collections, real inquiries and operational service workflows.

6. Conclusion

This paper discusses RAG, tool use and multi-step agentic retrieval in terms of retrieval control and considers their suitability for university libraries. Some evidence exists that iterative query, citation expansion and reranking can improve evidence coverage in some tasks. However, precision, reliability of evidence and operating cost do not automatically improve as workflows become more complex. The real value of agents is not to replace the expert judgment of the professional librarian but to structure the natural language query, external retrieval, screening of results, and iterative correction into a process that can be documented and reviewed.

DOI: 10.54254/2755-2721/2026.39322

According to the specific needs of paper discovery, reference service, resource navigation, and subject service, the degree of automation in the use of agents should be determined by university libraries. Some parameters need to be specified, including data sources, scope of retrieval, stopping conditions, and points of human intervention. The design of resource access, source verification, process logging, privacy protection, copyright compliance, and librarian review should be integrated together. Only after testing on real collections and service workflows should agents be trusted as professional assistants.

References

[1] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

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

[3] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model-based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

[4] Zhu, T. (2026). Construction and application of multi-agent systems for library subject services based on large language models [in Chinese]. *Digital Library Forum, 22*(2), 43–52. https://doi.org/10.3772/j.issn.1673-2286.2026.02.005

[5] Zhang, Y. (2026). AI virtual librarian service framework and implementation path in university libraries [in Chinese]. *Library Journal, 45*(1), 49–60. https://doi.org/10.13663/j.cnki.lj.2026.01.006

[6] Wang, Y. Y. (2025). Embedded knowledge services in university libraries based on large language models [in Chinese]. *Library Work and Study*, (10), 87–95.

[7] He, Y., Huang, G., Feng, P., Lin, Y., Zhang, Y., Li, H., & E, W. (2025). PaSa: An LLM agent for comprehensive academic paper search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11663–11679). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.572

[8] Shi, X., Li, Y., Kou, Q., Yu, L., Xie, J., & Zhou, H. (2025). SPAR: Scholar paper retrieval with LLM-based agents for enhanced academic search. arXiv. https://doi.org/10.48550/arXiv.2507.15245

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[10] Wang, X., Xia, T., Chen, J., Deng, X., Zu, C., Tian, Z., & Xiao, J. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 132576–132603). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/hash/efb741dfb0014e0c7a64104a70f236a4-Abstract-Datasets_and_Benchmarks_Track.html

[11] Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., & Li, Y. (2023). API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3102–3116). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.187

[12] Feng, J., Tao, C., Geng, X., Shen, T., Xu, C., Long, G., Zhao, D., & Jiang, D. (2024). Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9571–9583). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.517

[13] Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. (2024). Searching for best practices in retrieval-augmented generation. arXiv. https://doi.org/10.48550/arXiv.2407.01219

[14] Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic retrieval-augmented generation: A survey on agentic RAG. arXiv. https://doi.org/10.48550/arXiv.2501.09136

DOI: 10.54254/2755-2721/2026.39322

[15] Guo, L. M., Fu, Y. M., & Liu, Y. R. (2025). Application of retrieval-augmented thinking in library knowledge-mining services [in Chinese]. *Library and Information Service, 69*(21), 108–119. https://doi.org/10.13266/j.issn.0252-3116.2025.21.009

[16] Guo, L. M., & Fu, Y. M. (2024). Construction of a library large language model knowledge-service system integrating the ReAct pattern [in Chinese]. *Library Tribune, 44*(6), 61–70.

[17] Zhang, G. Z., Wang, Z. Y., Yang, F., Zhang, Y. L., & Wang, N. (2025). Deep integration of intelligent reference services and library operations [in Chinese]. *Library and Information Service, 69*(6), 33–45. https://doi.org/10.13266/j.issn.0252-3116.2025.06.003

[18] Fu, G. R., Wang, P. L., Wang, Y. Z., & Song, X. G. (2025). Construction and application of intelligent reference services in university libraries based on large language models: A case study of Shandong University Library [in Chinese]. *Library Journal, 44*(12), 31–40, 47. https://www.libraryjournal.com.cn/CN/Y2025/V44/I416/31

[19] Tong, Y. H., & Chen, J. L. (2025). Dual transformation amid the DeepSeek boom: Technological innovation in large models and reconstruction of university library service paradigms [in Chinese]. *Journal of Academic Libraries, 43*(1), 66–70. https://doi.org/10.16603/j.issn1002-1027.2025.01.008

[20] Wei, M. Y., Luo, K. M., Tang, M., et al. (2025). Embedding large language models in scenario-based smart library services [in Chinese]. *Library Theory and Practice*, (1), 83–94. https://doi.org/10.14064/j.cnki.issn1005-8214.20241029.001