The Social Science Review A Multidisciplinary Journal. May-June, 2026. Vol. 4. Issue 3. 229-233

Published by: Pather Dabi Educational Trust, (Regn No: IV-1402-00064/2023), Under Govt. of West Bengal, India

The Social Science Review
A Multidisciplinary Journal
ISSN: 2584-0789
(Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International e-Journal)
Homepage: www.tssreview.in

RETRIEVAL-AUGMENTED GENERATION FOR ACADEMIC LIBRARY REFERENCE AND DISCOVERY SERVICES: A FRAMEWORK FOR TRUSTWORTHY IMPLEMENTATION AND EVALUATION

Dheeraj Kumar

RESEARCH ARTICLE

Author Details:
Librarian, Veerangana Rani Durgavati Government Girls College Takhatpur, Bilaspur, Chhattisgarh, India

Corresponding Author:
Dheeraj Kumar

Abstract

Keyword-based discovery systems are increasingly mismatched with how patrons now express information needs, while general-purpose large language models (LLMs) answer fluently but fabricate sources and citations. Retrieval-augmented generation (RAG) couples a neural retriever with a generative model so that responses are grounded in an authoritative collection rather than the model's parametric memory alone. This paper synthesizes the emerging literature on RAG in academic libraries and argues that the central challenge for the field is not technical feasibility, which early prototypes have already demonstrated, but trustworthiness. Drawing on recent implementations and systematic reviews, the paper identifies four trust-related failure modes relevant to reference and discovery: residual hallucination, citation fabrication, conflict between the model's internal prior and retrieved evidence, and patron over-reliance. It then proposes a six-dimension evaluation framework, spanning retrieval quality, answer faithfulness, citation accuracy, calibrated abstention, equity and multilingual coverage, and privacy, intended to guide both system development and library assessment. The paper concludes that libraries are well positioned to lead responsible RAG deployment because their professional commitments to authoritative sourcing, patron privacy, and information literacy directly address the weaknesses of conversational AI.

DOI:
https://doi.org/10.70096/tssr.260403037

Keywords: *retrieval-augmented generation, academic libraries, information retrieval, large language models, discovery systems, AI evaluation*

Introduction

For decades, the catalog and the discovery layer have rested on the same basic mechanism: a patron supplies keywords, and the system returns documents whose metadata or full text match those terms. This model has served libraries well, but it assumes that users can translate an information need into the vocabulary of the index. In an environment shaped by conversational artificial intelligence, that assumption is increasingly fragile. Students who routinely pose full-sentence questions to chatbots arrive at the library catalog and find a tool that rewards a narrower, more mechanical kind of query. The gap between how people now express information needs and how discovery systems expect to receive them has become one of the more pressing usability problems in academic librarianship.

General-purpose large language models (LLMs) appear, at first glance, to close this gap. They accept natural-language questions and return fluent, synthesized answers. Yet they do so by predicting plausible text rather than by consulting a verified record, and the result is a well-documented tendency to fabricate facts and invent citations that look authoritative but do not exist. For a profession whose core value is connecting people to trustworthy information, an answer engine that optimizes for plausibility rather than truth is not an acceptable substitute for the catalog. The problem, then, is to combine the conversational accessibility of LLMs with the authoritative grounding of library collections.

Retrieval-augmented generation (RAG) is the architecture most often proposed to do exactly this. First formalized by Lewis et al. (2020), RAG pairs a retrieval component that searches an external knowledge source with a generative model that composes an answer conditioned on what was retrieved. Rather than relying solely on knowledge baked into model weights during training, the system grounds each response in documents drawn from a designated collection. The approach has moved quickly from research novelty to default practice; Brown et al. (2025), in a systematic review of the most-cited RAG literature, describe it as a method for grounding model output in up-to-date, non-parametric memory while preserving the broad linguistic competence

stored in the model. For libraries, the appeal is obvious: the non-parametric memory can be the institution's own repository, licensed databases, or curated reference collection.

This paper makes three contributions. First, it synthesizes the early but rapidly growing literature on RAG in academic library settings, distinguishing what has been demonstrated from what remains untested. Second, it argues that the decisive question for the field is no longer whether RAG can be built in a library context, since working prototypes already exist, but whether it can be made trustworthy enough for reference and discovery work. Third, it proposes a structured, six-dimension evaluation framework that libraries can use to assess RAG systems against their own professional standards rather than against generic benchmarks designed for commercial chatbots.

Technical Foundations of Retrieval-Augmented Generation

A RAG system has two cooperating parts. The retriever converts a user's query into a representation that can be matched against a stored collection, and the generator produces a natural-language response conditioned on both the query and the retrieved material. In contemporary implementations, retrieval is usually semantic rather than lexical. Documents are divided into passages and encoded as dense vector embeddings, numerical representations that place texts with similar meaning near one another in a high-dimensional space. The same encoding is applied to the incoming query, and the system retrieves the passages whose vectors lie closest to it. This is what allows RAG to answer a question phrased in everyday language even when the source documents use different terminology, a capability that traditional keyword matching lacks.

The retrieved passages are then inserted into the prompt given to the generative model, which synthesizes them into a coherent answer. Because the model is instructed to base its response on supplied evidence, it can, in principle, cite the specific sources it used and refrain from answering when nothing relevant is retrieved. Huang and Huang (2024), surveying retrieval-augmented text generation, document a steady proliferation of refinements to this basic loop: query rewriting to improve retrieval, re-ranking of candidate passages before generation, and iterative retrieval in which the system performs multiple rounds of search as it reasons through a complex question. Each refinement targets a known weakness, but each also adds latency and engineering complexity, a trade-off that matters acutely for libraries operating with constrained budgets and staff.

The way documents are divided before encoding, often called chunking, deserves particular emphasis in a library setting, because bibliographic and metadata records behave very differently from the continuous prose on which most RAG systems were first developed. A catalog record is short, highly structured, and dense with controlled vocabulary, whereas a digitized monograph is long and discursive. Chunk passages too finely and the system loses the context needed to interpret them; chunk them too coarsely and retrieval becomes imprecise and the generator is fed irrelevant text. Determining how to segment and encode the heterogeneous materials a library actually holds, from MARC records to full-text theses to finding aids, is an underexplored problem whose answer will differ from the defaults inherited from general-purpose implementations.

Two architectural choices shape how a library system behaves. The first is the boundary of the knowledge source: whether retrieval is restricted to the institutional repository, extended to licensed content, or allowed to reach the open web. A tighter boundary increases trust in provenance but narrows coverage. The second is whether the underlying model is used as delivered or adapted to the collection. Most library prototypes to date use general models without retraining, relying on retrieval alone to supply domain knowledge, which keeps cost low but leaves the system dependent on the quality of its retrieval step. These choices are not merely technical; they determine the values the system embodies, a point developed later in this paper.

Retrieval-Augmented Generation in the Library Context

Although RAG originated in computer science, its migration into library practice has been swift, and a small body of applied work now exists. Lund (2025) offers one of the first sustained treatments of RAG as library search infrastructure rather than as a general-purpose tool. He argues that RAG architectures provide a credible path from document-centric querying toward intent-aware, semantic knowledge discovery, and he surveys the technical components a library would need, including vector databases, embedding pipelines, and middleware to connect these to existing integrated library systems. Crucially, Lund frames the contribution of libraries as their capacity to supply authoritative, curated collections as the retrieval source, preserving the credibility that academic users expect even as the interface becomes conversational.

Concrete implementation has followed quickly. Xuan (2026) documents a technical case study at the University of Manitoba in which an institutional repository was connected to cloud-based AI services to support discovery. The prototype harvested repository metadata through a standard protocol, generated semantic embeddings, indexed them in a vector search service, and exposed both traditional keyword search and a generative, context-aware chat interface through a custom front end. The account is candid about engineering friction, from metadata harvesting errors to interface versioning conflicts, and it offers a reproducible roadmap for libraries attempting similar work. Its value lies less in any single result than in demonstrating that a functioning RAG discovery layer can be assembled from available components by a library team.

Parallel work in adjacent domains illustrates both promise and caution. Aytar et al. (2024) built a RAG framework for navigating academic literature in data science, combining bibliographic extraction, fine-tuned embeddings, and semantic chunking, and evaluated it with an automated assessment framework, reporting marked gains in the relevance of retrieved context. Afzal et al. (2024) tested several retrieval optimizations on university study-program data and introduced a structured evaluation device they call a RAG confusion matrix, underscoring that configuration choices substantially affect performance. On the collection-development side, Portillo and Carson (2025) evaluated four generative models for identifying gaps in a health sciences

collection and concluded that the models were not yet reliable as primary tools because of inaccuracies and hallucinations, though they served usefully as supplementary aids. Taken together, this literature establishes feasibility while repeatedly returning to a single theme: the systems work, but their outputs cannot yet be trusted without scrutiny.

What distinguishes the library application of RAG from a consumer chatbot is precisely the discipline imposed on its knowledge source. A general assistant retrieves from the open web, where provenance is uneven and authority is difficult to establish; a library system retrieves from a collection that has already been selected, described, and vetted by professionals. This inversion changes the value proposition. The library is not competing with commercial chatbots on breadth or fluency, where it will lose, but on trustworthiness of grounding, where its curated collections give it a genuine and defensible advantage. The implication is that investment should concentrate on the integrity of the retrieval source and the transparency of how answers are derived from it, rather than on matching the conversational polish of much larger commercial systems.

The Trustworthiness Problem: If feasibility is largely settled, trustworthiness is not. Reference and discovery are domains in which a confidently wrong answer is worse than no answer, because it can send a patron down an unproductive or misleading path while appearing authoritative. Four failure modes deserve particular attention.

Residual Hallucination: RAG reduces hallucination by grounding responses in retrieved evidence, but it does not eliminate it. When retrieval returns nothing relevant, or when the model overrides the retrieved passages with its own assumptions, it can still produce fabricated content. The collection-development findings of Portillo and Carson (2025) are a reminder that grounding is partial rather than absolute. In a discovery context, a system that occasionally invents a holding or misstates what a source contains erodes the trust that makes the library valuable in the first place.

Citation Fabrication: A specific and damaging form of error is the invented or misattributed citation. Patrons are likely to treat a library system as more authoritative than a consumer chatbot, which raises rather than lowers the stakes of a fabricated reference. A RAG system should, by design, cite only the passages it actually retrieved, yet ensuring that generated citations correspond faithfully to retrieved sources is a non-trivial engineering and evaluation problem that general benchmarks rarely measure directly.

Conflict Between Prior and Evidence: LLMs carry substantial knowledge in their parameters, and this internal prior can conflict with what the retriever supplies from the collection. When the two disagree, the system's behavior is consequential: a library system should generally privilege the authoritative retrieved source over the model's training-derived guess, yet models do not reliably do so. Brown et al. (2025) note that managing this tension is among the open methodological challenges in the RAG literature, and for libraries it is also a question of institutional values, since the whole point of grounding in a curated collection is to make that collection authoritative.

Patron Over-Reliance: A subtler risk is behavioral rather than technical. A fluent, synthesized answer may discourage patrons from examining sources themselves, weakening exactly the evaluative habits that information-literacy instruction aims to build. A discovery tool that answers too well, in the sense of foreclosing further inquiry, could undermine the educational mission of the academic library even when its individual answers are accurate. This connects RAG design directly to long-standing pedagogical concerns and suggests that good systems should be built to invite verification rather than to discourage it.

A Framework for Evaluating Library RAG Systems: General RAG benchmarks measure qualities that matter to commercial developers but capture only part of what libraries need. The framework proposed here organizes evaluation around six dimensions chosen to reflect professional library values. It is intended to be usable both by developers tuning a system and by librarians deciding whether a system is fit for public service.

Retrieval Quality: Because the generator can only be as good as what it is given, retrieval is the foundation. Evaluation should measure whether the passages returned for a representative set of real patron queries are actually relevant and sufficient to answer them, using established context-relevance and recall measures. The structured evaluation approaches reported by Aytar et al. (2024) and Afzal et al. (2024) offer practical starting points that libraries can adapt to their own collections and query logs.

Answer Faithfulness: Faithfulness asks whether the generated answer is actually supported by the retrieved passages, independent of whether those passages were the best available. A faithful but poorly retrieved answer and an unfaithful but well-retrieved one are different failures requiring different fixes, so the two dimensions must be measured separately. Faithfulness scoring, increasingly automated in recent assessment frameworks, should be a standard part of any library deployment review.

Citation Accuracy: Given the stakes of fabricated references in an academic setting, citation accuracy warrants its own dimension. Evaluation should verify that every citation in an answer corresponds to a genuinely retrieved source and that the cited source in fact supports the claim attached to it. This is a stricter standard than faithfulness alone and is, arguably, the dimension on which a library system most needs to outperform a consumer chatbot.

Calibrated Abstention: A trustworthy reference system must know the limits of its collection. When relevant material is not present, the appropriate response is to say so and to direct the patron to a human librarian or an alternative resource, not to improvise. Evaluation should therefore include queries deliberately outside the collection's scope and measure whether the system abstains appropriately rather than fabricating. Calibrated abstention is closely tied to reducing both hallucination and over-reliance.

Equity and Multilingual Coverage: Almost all current library RAG work is English-centric, yet academic communities are multilingual, and many institutions outside the dominant publishing languages are underserved. Evaluation should test performance across the languages a library's users actually speak and across query phrasings that reflect varied levels of academic preparation. Building and testing RAG over non-English collections remains a largely open and socially significant area, particularly for libraries in multilingual regions, where equitable discovery cannot be assumed from English-language results.

Privacy: Libraries hold patron-privacy as a core professional commitment, but cloud-based RAG architectures route queries, which can themselves be sensitive, through third-party services. Evaluation should account for where query data travels, whether it is retained, and whether a more privacy-preserving architecture, including locally hosted models, could meet the need. A discovery tool that compromises patron confidentiality fails on a dimension that no accuracy metric will capture.

These six dimensions are not independent. Strengthening abstention, for instance, tends to reduce both hallucination and over-reliance, and a privacy-driven choice to run models locally will constrain which retrieval and generation techniques are available. The framework's purpose is not to yield a single score but to make the trade-offs visible so that libraries can weigh them deliberately rather than inheriting the priorities embedded in tools built for other purposes.

Governance, Ethics, and Equity

Evaluation shades into governance, because some of the most consequential questions about library RAG are not about accuracy at all. Copyright and licensing are immediate concerns: when a system retrieves from licensed databases and reproduces passages within a generated answer, it may exceed the terms under which the content was acquired, and the legal contours of this practice remain unsettled. Libraries will need clear policies on which sources a RAG system may draw from and how much retrieved text it may surface.

Transparency is a second governance concern. As public institutions, academic libraries are accountable to their communities in ways that commercial vendors are not, which implies an obligation to disclose when answers are AI-generated, to make the basis of those answers inspectable, and to log system behavior for review. A discovery layer whose reasoning is opaque sits uneasily with the library's role as a trustworthy public resource.

Finally, equity runs through the entire enterprise. The most capable RAG implementations described in the literature rely on commercial cloud services that carry recurring costs many libraries cannot sustain, raising the prospect that conversational discovery becomes a privilege of well-funded institutions. Investigating low-cost, open-source RAG stacks, built on openly available embedding models, self-hosted vector stores, and open-weight language models, is therefore not merely a technical exercise but a question of access. Combined with the multilingual gap noted above, this points to a research and practice agenda in which the goal is not only systems that work, but systems that work for institutions and communities currently at the margins of the technology.

Accountability also has a practical, operational dimension that libraries are well equipped to handle. Maintaining logs of queries and the sources retrieved in response, subject to privacy safeguards, allows a library to audit system behavior, identify recurring failure patterns, and correct them over time. This kind of continuous oversight is consistent with how libraries already steward their collections and services, and it distinguishes a responsibly operated institutional tool from a black-box commercial product. Treating a RAG system as an ongoing service to be monitored, rather than a finished product to be installed, is likely the posture most compatible with professional norms.

Implications for Practice and Future Research

For practitioners, the immediate lesson is to treat RAG as an augmentation of, not a replacement for, existing discovery and reference services, and to deploy it with explicit evaluation against the dimensions above before it reaches patrons. Pilot projects should pair any conversational interface with clear signposting toward human assistance and toward the underlying sources, preserving the verification habits that information-literacy instruction depends on.

For researchers, several questions are both tractable and pressing. A focused empirical study could measure citation-fabrication rates across retrieval configurations on a real collection, an investigation that requires a query set and scoring rubric rather than large infrastructure. Comparative work could test whether conversational discovery complements or cannibalizes traditional catalog use. And design-oriented research could ask how to build systems that encourage source verification rather than discourage it. The multilingual and low-cost directions, meanwhile, reward sustained projects tied to a specific collection and community, and would extend the literature beyond its current English-centric, well-resourced frame.

Implementation also raises questions of staffing and skills that the technical literature tends to overlook. Building and maintaining a RAG discovery layer draws on competencies, including vector search, embedding pipelines, and prompt design, that are not yet standard in library technology teams, and sustaining such a system requires ongoing attention rather than a one-time deployment. Libraries considering this work will need to weigh whether to develop these capacities internally, collaborate across institutions to share the burden, or rely on vendors, each option carrying different implications for cost, control, and the privacy concerns raised earlier. The choice is as much organizational as technical, and it deserves explicit deliberation rather than being settled by default through a procurement decision.

Conclusion

Retrieval-augmented generation offers academic libraries a credible way to meet patrons who now expect to ask questions in natural language, without abandoning the authoritative grounding that distinguishes a library from a consumer chatbot. The early literature shows that such systems can be built; the harder and more important work is making them trustworthy. The six-dimension framework proposed here, spanning retrieval quality, faithfulness, citation accuracy, calibrated abstention, equity and multilingual coverage, and privacy, is intended to help libraries hold these systems to their own standards rather than to standards imported from commercial AI. Libraries are, in fact, unusually well placed to lead responsible deployment, because the very weaknesses of conversational AI, ungrounded claims, opaque sourcing, indifference to privacy, and the displacement of critical inquiry, map directly onto commitments the profession has held for a long time. Realizing that potential will require treating evaluation, governance, and equity as central rather than incidental to the design of the next generation of discovery tools.

Acknowledgment: No
Author's Contribution: *Dheeraj Kumar:* Data Collection, Literature Review, Methodology, Analysis, Drafting, Referencing;
Funding: No
Declaration: The author has given consent for the publication.
Competing Interest: No

References

1. Afzal, A., Vladika, J., Fazlija, G., Staradubets, A., & Matthes, F. (2024). *Towards optimizing a retrieval augmented generation using large language model on academic data*. arXiv. https://arxiv.org/abs/2411.08438
2. Aytar, A. Y., Kaya, K., & Kilic, K. (2024). *A retrieval-augmented generation framework for academic literature navigation in data science*. arXiv. https://arxiv.org/abs/2412.15404
3. Brown, A., Roman, M., & Devereux, B. (2025). *A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges*. arXiv. https://arxiv.org/abs/2508.06401
4. Huang, Y., & Huang, J. (2024). *A survey on retrieval-augmented text generation for large language models*. arXiv. https://arxiv.org/abs/2404.10981
5. 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. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates.
6. Lund, B. (2025). Prospects of retrieval augmented generation (RAG) for academic library search and retrieval. *Information Technology and Libraries, 44*(2). https://ital.corejournals.org/index.php/ital/article/view/17361
7. Portillo, I., & Carson, D. (2025). Making the most of artificial intelligence and large language models to support collection development in health sciences libraries. *Journal of the Medical Library Association, 113*(1). https://doi.org/10.5195/jmla.2025.2079
8. Xuan, W. (2026). Implementing retrieval-augmented generation for academic libraries: A technical case study using Azure AI. *International Journal of Librarianship, 11*(1), 151–166. https://doi.org/10.23974/ijol.2026.vol11.1.601