

Personalized Academic Information Retrieval Using User Profiles and Large Language Model Agents

Felix R. Mendez

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.

felix.r.mendez@oregonstate.edu

Zhouzhou Xu

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.

zhouzhou619@unr.edu

Abstract

Academic information retrieval has long struggled with a fundamental tension between generic relevance and individual scholarly context. The rapid growth of digital libraries, preprint repositories, and interdisciplinary publication venues has intensified this tension by fragmenting the search landscape and increasing the ambiguity of short queries. This paper examines the role of large language model agents in enabling personalized academic information retrieval through the construction, maintenance, and operational use of rich user profiles. Rather than treating personalization as a lightweight ranking adjustment, the paper argues that effective academic search requires a system-level integration of explicit and implicit user signals, contextual query understanding, retrieval orchestration, explanation generation, and long-term memory. We analyze architectural configurations in which large language model agents act as mediators between user profiles and heterogeneous retrieval backends, including sparse lexical indexes, dense vector stores, and external academic databases. The discussion emphasizes structural trade-offs related to latency, provenance, interpretability, privacy, fairness, and sustainability. We also consider governance challenges that arise when user profiles are inferred from reading behavior, writing trajectories, and citation practices. The paper contributes a conceptual framework for understanding personalized academic retrieval as a socio-technical infrastructure problem rather than a purely algorithmic optimization task. It concludes with directions for robust evaluation, policy-aware design, and sustainable deployment in university libraries, research institutions, and scholarly search platforms.

Keywords

personalized information retrieval; large language models; user profiling; retrieval-augmented generation; academic search; algorithmic governance; scholarly infrastructure.

1. Introduction

Academic information retrieval has traditionally been understood as a process of matching query representations to document representations under conditions of uncertainty and partial knowledge. Early work emphasized that information retrieval and information filtering are closely related but distinct activities, with retrieval addressing ad hoc information needs and filtering addressing stable long-term interests [1]. This distinction remains important for personalized academic search because researchers often have both temporary query-level

goals and durable thematic commitments that evolve over years. The stratified model of information retrieval interaction further highlighted that users move through multiple cognitive and situational layers when formulating queries, judging relevance, and interpreting retrieved results [2]. These foundational insights suggest that personalization cannot be reduced to adding a few user-specific terms to a query. Instead, it requires continuous modeling of the user, the task, and the scholarly context in which information will be used.

The modern academic information environment amplifies these challenges. Researchers must navigate disciplinary databases, institutional repositories, preprint servers, publisher platforms, and increasingly heterogeneous scholarly artifacts such as datasets, code repositories, and registered reports. Query logs from academic systems reveal substantial variability in how scholars express similar information needs, and conventional evaluation measures such as cumulative gain were developed to capture graded relevance judgments across multi-level usefulness rather than binary topical match [3]. In this setting, a personalized retrieval system can improve both precision and user satisfaction by aligning results with the researcher’s prior knowledge, methodological preferences, and current project trajectory. However, achieving this alignment requires more than static filtering. It demands a dynamic infrastructure that can interpret intent, negotiate between multiple retrieval mechanisms, and explain why particular documents were selected.

Large language model agents introduce new possibilities for this infrastructure. Unlike conventional recommender systems that rely primarily on historical interaction patterns, large language model agents can reason over natural language descriptions of research interests, synthesize fragments from multiple sources, and produce explanations that are legible to scholars. At the same time, their integration into academic retrieval raises significant questions about provenance, controllability, and long-term adaptation. This paper adopts a systems perspective and examines how user profiles and large language model agents can be combined to create personalized academic information retrieval environments that are not only effective but also governable, fair, and sustainable.

2. Foundations and Motivations for Personalized Academic Retrieval

The foundations of personalized academic retrieval lie in classical information retrieval, user modeling, and recommender systems. Classical information retrieval frameworks established that relevance is a relation between a user’s information need and a document, yet this relation is mediated by uncertain representations of both need and content [4]. In academic search, the user may be a doctoral student entering a new field, an established researcher testing an interdisciplinary hypothesis, or a reviewer seeking methodological precedents. These roles imply different expectations for document novelty, authority, and technical depth. Personalization therefore involves understanding the user’s role, discipline, and epistemic goals as part of the relevance judgment process.

Recommender systems research has long addressed the challenge of aligning recommendations with long-term preferences. A broad survey of recommender system evolution showed that personalized systems often combine collaborative signals, content features, and hybrid methods to capture user preferences under sparse feedback conditions [5]. Matrix factorization techniques have been particularly influential because they can uncover latent factors that explain observed user-item interactions without requiring exhaustive feature engineering [6]. Although academic search is not identical to recommendation, it shares the need to infer preferences from sparse and noisy behavioral signals, including downloads, citations, saved references, and time spent reading abstracts.

Neural information retrieval has further transformed how relevance can be modeled. Pre-trained transformer architectures such as BERT enabled contextual representations of queries and documents, leading to substantial gains in tasks that require deep semantic matching [7]. The underlying self-attention mechanism introduced by the transformer made it possible to represent long-range dependencies in text more effectively than previous recurrent or convolutional approaches [8]. These advances matter for academic retrieval because scholarly language is dense, terminologically specialized, and often ambiguous across disciplines. A user profile that captures the semantic space in which a researcher operates can help disambiguate terms that have different meanings in different fields, such as ontology, resilience, or normalization.

The shift toward neural and contextual retrieval also motivates a rethinking of academic search as an interactive process. Rather than treating personalization as a post hoc re-ranking step, modern systems can maintain rich user models that influence query expansion, document encoding, result diversity, and explanation. This requires a system architecture capable of coordinating multiple components with different update cycles, failure modes, and computational costs. The following sections develop such an architecture by focusing on the integration of user profiles and large language model agents.

3. System Architecture for User Profiles and Large Language Model Agents

A personalized academic information retrieval system can be organized around a modular architecture in which user profiles, retrieval backends, and large language model agents interact through well-defined interfaces. At the center of this architecture is an orchestration layer that translates a user's expressed information need into a set of retrieval actions and then synthesizes the results into a coherent response. Modern sequence-to-sequence language models have demonstrated that many language understanding tasks can be reframed as text generation problems, which is particularly useful for query rewriting and intent normalization [9]. This capability allows an agent to convert a brief query such as causal inference in observational studies into a more precise representation that accounts for the user's disciplinary background and methodological preferences.

Retrieval in such a system is unlikely to rely on a single index. Academic collections may include sparse lexical indexes for exact terminology, dense passage retrievers for semantic similarity, and metadata filters for author, venue, and publication date. Dense passage retrieval has proven effective in open-domain question answering by encoding queries and passages into a shared vector space and retrieving nearest neighbors efficiently [10]. Yet dense retrieval can miss exact matches for rare technical terms that matter greatly in scholarly search. Consequently, a hybrid retrieval strategy is often necessary. Large language model agents can decide when to issue lexical queries, when to perform dense retrieval, and when to combine both, based on the query type and the user profile.

The integration of retrieval with generation has been formalized through retrieval-augmented generation, in which a language model conditions its output on retrieved passages to improve factual grounding [11]. In academic retrieval, this approach can support grounded summaries of the retrieved literature, citation-aware responses, and explanations that connect a recommended paper to the user's prior work. A recent survey of retrieval-augmented generation has highlighted its rapid adoption and its remaining challenges in retrieval quality, latency, and evaluation [12]. These challenges are particularly pronounced in academic domains because retrieved documents are often long, densely cited, and conceptually complex. The architecture must therefore support document chunking strategies, citation boundary

detection, and selective conditioning on abstracts or full texts depending on the user's need and the agent's confidence.

4. User Profile Modeling: Signals, Semantics, and Lifecycle

User profile modeling is the foundation of personalized academic retrieval. Profiles may be constructed from explicit information such as self-reported research interests, departmental affiliation, and publication records, as well as implicit signals including query history, click behavior, saved papers, and time spent reading. Early work in unsupervised language modeling demonstrated that large corpora can yield rich contextual representations that capture syntactic and semantic regularities without task-specific supervision [13]. These representations can be used to encode a scholar's publications, reading lists, and query history into a semantic profile that evolves as the scholar's interests shift.

A robust user profile model must account for multiple temporal scales. Short-term signals capture immediate task context, such as preparing a lecture or reviewing related work for a paper. Medium-term signals reflect active research projects and may persist for months or years. Long-term signals represent enduring intellectual commitments, methodological preferences, and interdisciplinary orientations. The profile lifecycle therefore requires mechanisms for decay, consolidation, and conflict resolution. For example, a researcher who previously worked on symbolic reasoning may shift toward neural methods, but the profile should not simply discard the older interest because it may remain relevant for historical context or future interdisciplinary work.

The semantic representation of user profiles also raises questions about abstraction and interpretability. Profiles expressed as continuous vectors can support efficient similarity computation but may be difficult for users to inspect or challenge. Profiles expressed as structured knowledge graphs can be more transparent but may require more expensive maintenance. A hybrid approach can represent stable attributes explicitly while using dense representations for exploratory interest modeling. The system must also support profile correction, because inferred interests can be wrong, outdated, or sensitive. Without user-facing controls, personalization may reinforce narrow search behavior and reduce serendipity, which is particularly undesirable in academic contexts where interdisciplinary discovery is valuable.

5. Large Language Model Agents as Retrieval Orchestrators

Large language model agents serve multiple roles in personalized academic retrieval. They can interpret ambiguous queries, infer task intent, select retrieval strategies, decompose complex information needs into subqueries, and synthesize results. The few-shot learning capabilities of large language models opened the possibility of adapting models to new tasks through natural language instructions and examples rather than extensive fine-tuning [14]. Instruction-tuned variants have further improved the ability of models to follow user goals and produce helpful responses in interactive settings [15]. These properties make large language model agents well suited to mediate between the user and the retrieval system, especially when the user's need is not fully articulated.

However, an agent that merely rewrites queries and summarizes results is insufficient for academic retrieval. The agent must also maintain a form of memory that reflects the user's evolving interests, previous queries, and stated preferences. Studies of user memory systems for query matching and intent modeling indicate that persistent memory can significantly enhance search efficiency by allowing the system to resolve ambiguous terms and anticipate

related information needs [16]. In academic contexts, such memory can help distinguish between a query about transformers in electrical engineering and transformers in natural language processing, or between a researcher seeking introductory material and one seeking state-of-the-art methods.

The agent's orchestration role also includes strategic retrieval. For a query that asks for recent methodological advances, the agent may prioritize venue freshness and citation velocity. For a query that asks for foundational theory, the agent may prioritize highly cited older works and authoritative review articles. The agent can also determine when to search multiple databases, when to query the user's personal library first, and when to include resources outside the user's stated discipline to encourage interdisciplinary discovery. These decisions must be grounded in both the user profile and a model of the underlying academic information landscape.

In addition, the agent must communicate its reasoning without overwhelming the scholar. Academic users are often capable of interpreting complex evidence, but they are also time-poor and skeptical of opaque recommendations. Explanations that reference specific sections of a paper, relate a recommendation to the user's prior publication, or indicate how a result supports a current project can increase trust and allow users to calibrate their reliance on the system. This interactive dimension connects personalized retrieval to broader concerns in human-AI interaction.

6. Retrieval, Ranking, and Explanation Under Personalization

The retrieval and ranking stage determines which documents surface for a given query and how they are ordered. Personalization at this stage can take many forms, from query expansion based on profile terms to personalized reranking of a candidate set. A central challenge is balancing personal relevance against scholarly quality, novelty, and diversity. A ranking model that overweights the user's prior publications may produce an echo chamber of familiar authors and venues, while a model that ignores the profile may fail to distinguish between similarly relevant documents. The system must therefore encode multiple ranking objectives and expose enough control for users to adjust the trade-off.

Explanation generation is closely connected to ranking. When a system recommends a paper, it should provide a rationale that is both accurate and useful. The rationale might identify overlapping concepts between the paper and the user's profile, highlight methodological similarity, or note that the paper has been highly cited by authors the user follows. Human-AI interaction guidelines emphasize that systems should make their capabilities and limitations clear, support efficient correction, and provide contextually appropriate explanations [17]. In academic retrieval, this means avoiding misleading claims about a paper's contribution and clearly indicating when a recommendation is based on inferred rather than explicit interests.

The integration of user profiles with ranking also raises questions about evaluation. A personalized result list may be difficult to evaluate using standard relevance judgments because relevance is partly conditioned on the individual user's context. Simulation-based evaluation, user studies, and longitudinal field trials can provide complementary evidence. The system should log not only which documents were clicked but also how users modified queries, whether they saved or exported references, and whether they rated explanations as helpful. These signals can then be used to refine both the user profile and the retrieval policy.

7. Governance, Privacy, Fairness, and Policy Implications

Personalized academic retrieval systems operate on sensitive intellectual data. A researcher’s query history can reveal unpublished ideas, grant proposals, health conditions, or political interests. The construction of user profiles therefore creates serious privacy risks if data are shared across institutions, used for surveillance, or retained without clear limits. Fairness and accountability scholars have argued that sociotechnical abstraction can hide the ways in which machine learning systems distribute benefits and harms unevenly [18]. In academic search, personalization may favor well-resourced institutions with richer behavioral data, further entrenching existing inequalities in scholarly visibility and access.

The ethics of algorithmic systems provides a broader lens for assessing these systems. Algorithmic decision-making raises concerns about transparency, responsibility, and the potential for unintended consequences when outputs influence human choices [19]. Personalized academic retrieval can shape what researchers read, cite, and pursue, thereby influencing the direction of scientific inquiry. If profile inference amplifies popular topics and suppresses long-tail scholarship, the system may narrow the collective research agenda. Governance mechanisms should therefore include algorithmic impact assessments, data minimization requirements, and user rights to inspect and delete inferred profiles.

Language models themselves introduce additional policy concerns. Large models may reproduce biases present in training corpora, including geographic, gender, and disciplinary biases in scholarly communication. The debate around stochastic parrots has highlighted the risks of treating text generation as a substitute for grounded understanding [20]. In academic retrieval, a language model agent that generates summaries or explanations must be constrained by retrieved evidence and clearly distinguish between statements drawn from sources and statements inferred from the model’s prior knowledge. Without such constraints, the system may fabricate citations, misattribute ideas, or overstate consensus.

Institutional governance should be designed to preserve scholarly autonomy. Users should be able to select what parts of their profile are used for personalization, opt out of behavioral tracking, and control whether their query history is stored. University libraries and academic publishers may need shared agreements on data provenance, model access, and audit requirements. Policy frameworks should also address the environmental and economic sustainability of operating large language model agents at scale.

8. Infrastructure, Scalability, and Operational Trade-offs

Deploying personalized academic retrieval with large language model agents involves difficult infrastructure trade-offs. Large language model inference is computationally expensive, and academic search systems may face highly variable query loads around deadlines, conference cycles, and semesters. Caching, model distillation, and selective invocation can reduce cost, but each introduces trade-offs in freshness and personalization quality. Code-oriented large language models have shown that specialized capabilities can be developed through targeted training, yet these capabilities still require significant compute and careful evaluation [21]. The same logic applies to domain-specific academic retrieval models.

Hybrid retrieval over large academic collections demands efficient indexing and search strategies. Dense retrieval systems can produce high-quality semantic matches but require substantial storage and compute for embedding creation and nearest neighbor search. Approaches such as contextualized late interaction have attempted to combine the expressiveness of transformer-based relevance matching with the efficiency of precomputed

token representations [22]. These trade-offs become more acute when the system must support personalized indexes that differ across users, because maintaining a separate vector space per user is infeasible at scale. A practical design may use a shared retrieval index with personalized query rewriting and reranking rather than fully personalized indexes.

Operational robustness also depends on the graceful degradation of personalized features. If a user profile is sparse or newly created, the system should fall back to non-personalized retrieval without misleading the user. If a large language model agent cannot access an external database due to an outage, the system should still return results from available sources and clearly indicate any limitations. This requires careful design of service dependencies, timeout policies, and fallback paths. It also requires observability mechanisms that can detect when personalization is degrading retrieval quality rather than improving it.

Sustainability considerations extend beyond energy consumption to include the maintenance burden of user profiles, model updates, and evaluation pipelines. Academic institutions may lack the engineering resources of commercial search providers. Consequently, the design should favor modular components that can be operated independently and replaced without system-wide disruption. Shared infrastructure consortia, open models, and standardized interfaces for academic metadata can reduce duplication and improve long-term viability.

9. Evaluation, Robustness, and Sustainability

Evaluating personalized academic retrieval requires moving beyond static benchmark scores. A retrieval system may achieve strong results on a generic academic query set while performing poorly for users whose profiles are underrepresented or whose disciplines use terminology in unusual ways. Large language model surveys have emphasized the importance of task-specific evaluation and highlighted the variability of model performance across domains [23]. Personalized systems should therefore be evaluated on multiple dimensions, including relevance, novelty, diversity, explanation quality, and user satisfaction. Longitudinal evaluation is particularly important because the value of personalization accrues over time as the user profile stabilizes and the system learns to anticipate recurring needs.

Robustness is another central concern. User profiles can be noisy, incomplete, and subject to rapid change. A scholar may move into a new field, collaborate across disciplines, or adopt a critical stance toward a previously accepted methodology. The system should detect when the profile no longer explains observed behavior and trigger re-inference or user confirmation. It should also resist manipulation by users or external actors who might attempt to distort retrieval rankings by exploiting profile signals. The broader risks of foundation models, including their opacity, bias, and concentration of power, apply directly to personalized academic search [24].

Sustainability includes not only environmental and financial dimensions but also institutional and epistemic sustainability. Personalized academic retrieval systems should support long-term scholarly practices such as literature review, citation tracing, and research synthesis, rather than optimizing only for immediate click satisfaction. This requires collaboration between computer scientists, librarians, publishers, and policy scholars to develop shared evaluation frameworks and infrastructure standards. It also requires continued attention to open science principles, ensuring that personalization does not become a proprietary layer that restricts access to the scholarly record.

10. Conclusion

Personalized academic information retrieval is evolving from a set of ranking heuristics into a complex socio-technical infrastructure. Large language model agents, when combined with carefully governed user profiles, offer the potential to interpret scholarly intent, orchestrate hybrid retrieval, and explain recommendations in ways that were previously difficult to achieve. However, realizing this potential requires attention to architecture, privacy, fairness, operational robustness, and long-term sustainability. The system must balance the immediate benefits of personalization against the risks of narrowed discovery, biased inference, and opaque decision-making.

This paper has argued that a systems perspective is essential. Personalized retrieval should be understood as an ongoing coordination problem among users, institutions, models, and scholarly resources. Future work should develop open evaluation protocols that account for longitudinal user outcomes, design governance mechanisms that protect scholarly autonomy, and build modular infrastructure that can be maintained by academic institutions. By treating personalization as a governance problem rather than only an algorithmic problem, the research community can create retrieval systems that are both intelligent and accountable.

References

1. Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12), 29–38.
2. Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the ASIS Annual Meeting*, 34, 313–327.
3. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446.
4. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
5. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
6. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37.
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
10. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781.

11. 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. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
12. 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 preprint arXiv:2312.10997*.
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.
14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
16. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
17. Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.
18. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.
19. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
20. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
21. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
22. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.
23. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
24. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.