Preference-Aware Neural Reranking with Collaborative Large Language Model Agents

Teuvo Nuutinen
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
nuutinenteuvo@colostate.edu

Kamil Junk
Department of Computer Science, University of Houston, Houston, TX, USA.
kamiljunk506@uh.edu

Abstract

Neural reranking has become a central mechanism for improving search, recommendation, and retrieval-augmented generation systems by reordering candidate items according to learned relevance signals. Recent advances in large language models have expanded reranking from single-pass scoring toward multi-agent collaboration, where different agents assume specialized roles such as query interpretation, evidence assessment, preference elicitation, and final arbitration. This paper examines preference-aware neural reranking through the lens of collaborative large language model agents. It argues that the next generation of reranking systems should not be understood merely as model architectures but as socio-technical infrastructures that coordinate user preferences, institutional constraints, computational resources, and governance obligations. The analysis develops a system-level framework covering architectural decomposition, agent coordination, preference representation, fairness, robustness, deployment sustainability, and policy implications. It considers how preference-aware reranking can improve personalization and contextual relevance while introducing new risks related to manipulation, opacity, feedback loops, and uneven access. The paper also discusses evaluation challenges, including the need for heterogeneous benchmarks, auditable logs, model documentation, and lifecycle monitoring. The central claim is that collaborative large language model agents can make reranking more adaptive and pluralistic, but only if their coordination protocols, preference models, and deployment environments are designed with explicit attention to accountability, resource asymmetry, and long-term maintainability. The paper concludes that preference-aware reranking should be treated as a governed infrastructure rather than a narrow ranking optimization problem.

Keywords

neural reranking, large language models, collaborative agents, preference modeling, information retrieval, socio-technical systems, governance, fairness, robustness, sustainability.

1. Introduction

Information retrieval systems increasingly depend on reranking to transform broad candidate sets into ordered results that appear relevant, useful, and contextually appropriate. Early retrieval pipelines often relied on lexical matching and statistical weighting, but neural architectures changed the design space by learning dense representations and task-specific relevance functions [1]. Transformer-based encoders and cross-attention rerankers demonstrated that deeper semantic comparison could substantially improve ranking quality

[2]. With the emergence of large language models, reranking has moved further toward generative and reasoning-oriented formulations in which the system can interpret nuanced queries, synthesize evidence, and produce justifications [3]. This shift has also encouraged retrieval-augmented generation, where ranking quality directly affects downstream factual grounding and user trust [4]. Yet these developments are often described as model improvements, while the surrounding system-level consequences remain underexamined.

Preference-aware reranking adds another layer of complexity because relevance is not a stable, universal property. Users, institutions, and communities hold different notions of usefulness, diversity, novelty, credibility, safety, and serendipity. A preference-aware system must therefore represent multiple, sometimes conflicting, evaluative criteria and decide how those criteria are combined. Collaborative large language model agents offer one response to this challenge by dividing ranking labor among specialized agents that can deliberate, critique, and revise one another's judgments. A recent multi-role collaborative reranking workflow illustrates how such decomposition can be operationalized in practical retrieval settings [17]. However, collaborative agents do not eliminate the underlying tension between personalization and public accountability; they redistribute it across prompts, coordination protocols, memory stores, and human oversight arrangements.

This paper argues that preference-aware neural reranking with collaborative large language model agents should be approached as a socio-technical infrastructure. The relevant questions are not only whether an agent ensemble improves ranking metrics, but also who defines preferences, how conflicts are resolved, what computational resources are consumed, how failures are detected, and how affected parties can contest outcomes. Drawing on information retrieval foundations, foundation model research, safety scholarship, and machine learning operations literature, the paper develops a structured analysis of architecture, coordination, fairness, deployment, robustness, and governance [8], [10], [18]. The aim is to provide a forward-looking account that is useful for researchers, system architects, policy analysts, and institutional operators.

2. Conceptual Foundations and Problem Setting

Reranking sits between candidate generation and final presentation. In classical retrieval, scoring functions such as BM25 provide efficient and interpretable ranking based on term statistics and document structure [7]. Dense retrieval methods later improved semantic matching by embedding queries and documents into shared vector spaces, enabling retrieval beyond exact lexical overlap [5]. Neural rerankers then applied deeper interaction models to a smaller candidate set, often achieving stronger relevance judgments at higher computational cost [6]. The resulting pipeline is hierarchical: recall-oriented retrieval feeds precision-oriented reranking, which in turn supports user-facing consumption or downstream generation. This hierarchy remains relevant when large language model agents are introduced, because agents are expensive and must be deployed selectively rather than universally.

The problem setting for preference-aware reranking is complicated by the fact that preferences are multi-dimensional. A user may prefer recent documents but also require authoritative sources, diverse viewpoints, and low cognitive load. An institution may prioritize compliance, safety, and reputational risk, while a platform may prioritize engagement and retention. A public-interest system may prioritize accessibility, pluralism, and democratic accountability. Collaborative large language model agents can represent these dimensions through role differentiation, but representation alone does not resolve normative conflict. The system must still contain a mechanism for aggregation, arbitration, and appeal.

Benchmarks such as BEIR have shown that zero-shot retrieval and reranking performance varies substantially across domains, highlighting the fragility of assuming a single relevance notion [20]. Preference-aware systems intensify this issue because they introduce user- and context-specific signals that may not be captured by static test collections. A ranking that is excellent for one population can be harmful or irrelevant for another. Therefore, the conceptual foundation for preference-aware reranking must include not only relevance estimation but also preference provenance, conflict resolution, and distributional impact. The system-level question is how to make ranking decisions that are adaptive without becoming arbitrary, personalized without becoming opaque, and efficient without externalizing hidden costs.

3. Architecture of Preference-Aware Neural Reranking

A preference-aware reranking architecture typically contains several layers. The candidate generation layer retrieves a broad set of items from lexical indexes, dense indexes, or hybrid sources [5], [7]. The representation layer encodes queries, items, user context, and institutional constraints. The agent layer assigns specialized reasoning tasks to large language model agents, such as query clarification, evidence verification, diversity promotion, or preference matching [3], [18]. The arbitration layer combines agent outputs into a final order. The governance layer logs decisions, monitors drift, and supports audit or appeal. Each layer introduces trade-offs between latency, cost, transparency, and adaptability.

Collaborative agents are attractive because a single monolithic ranker may struggle to satisfy competing objectives. One agent can focus on topical relevance, another on source credibility, another on preference alignment, and another on safety or legal compliance. This division of labor resembles ensemble methods but differs in that agents may communicate in natural language, produce rationales, and revise their positions. Retrieval-augmented generation architectures provide a useful reference because they also coordinate retrieval, evidence selection, and generation [4], [19]. In reranking, however, the final output is an ordered list rather than a textual answer, so the system must translate deliberative discourse into a stable ranking decision. That translation step is a major source of design risk because small changes in aggregation rules can produce large changes in exposure and user experience.

A robust architecture should distinguish between preferences that are explicitly declared, implicitly inferred, and institutionally imposed. Explicit preferences may come from user settings or queries. Implicit preferences may be inferred from behavior, dwell time, or prior interactions. Institutional preferences may come from policy, law, or organizational mission. Collaborative agents can treat these categories differently, but they must also preserve provenance so that a ranking can be explained and contested. Without provenance, preference-aware reranking becomes a mechanism for hidden influence. With provenance, it can become a form of accountable personalization. The architecture therefore needs metadata standards, versioned preference profiles, and clear boundaries between short-term contextual adaptation and long-term user modeling.

4. Collaborative Large Language Model Agents and Coordination

Coordination among large language model agents is not merely a technical orchestration problem. It is a governance problem because coordination protocols determine whose judgment prevails, how uncertainty is represented, and when human review is required. A multi-role workflow can assign agents to critique, rank, and consolidate, as demonstrated in recent collaborative reranking research [17]. Such designs can improve robustness by

exposing disagreements rather than hiding them. For example, if one agent emphasizes topical match and another emphasizes source reliability, the system can surface the tension instead of collapsing it into a single opaque score. This can be valuable in high-stakes domains such as health, law, education, and public information.

However, collaboration also introduces failure modes. Agents may share correlated biases if they are built on similar foundation models or trained on overlapping data [9], [10]. They may converge prematurely through social pressure embedded in prompts or debate protocols. They may generate plausible rationales that do not faithfully reflect the underlying ranking logic. They may consume substantial compute, creating latency and environmental costs. They may also amplify majority preferences while marginalizing minority perspectives. Coordination protocols must therefore include diversity checks, uncertainty thresholds, conflict escalation, and periodic revalidation against external evidence.

A practical coordination design might separate roles into retrieval interpretation, candidate assessment, preference alignment, risk review, and final arbitration. Yet the paper avoids prescribing a fixed algorithm because the appropriate division depends on domain, scale, and governance context. The key insight is that agent roles should be explicit, auditable, and revisable. Each role should have documented objectives, known limitations, and escalation paths. The system should maintain a decision trace that records which agents contributed, what evidence was used, and how conflicts were resolved. This trace is not only useful for debugging but also for regulatory compliance, user redress, and public accountability. In this sense, collaborative agents shift reranking from a score-computation problem to a deliberative infrastructure problem.

5. Preference Modeling, Fairness, and User Representation

Preference modeling is often treated as a technical prediction task, but it is fundamentally a representational and political act. Decisions about which signals count as preference, how they are weighted, and how long they persist determine which users are served well and which are rendered invisible. Large language models can infer preferences from natural language, but such inference is probabilistic and context-sensitive [11]. It may reproduce stereotypes, overgeneralize from limited evidence, or confuse stated preferences with inferred behavior. A preference-aware reranker must therefore distinguish between preference prediction and preference authorization. Users should have meaningful control over the preferences that shape their results, especially when ranking affects access to opportunities, information, or services.

Fairness in reranking cannot be reduced to equality of exposure across groups. It also involves procedural fairness, informational fairness, and allocative fairness. Procedural fairness asks whether ranking decisions follow transparent and contestable rules. Informational fairness asks whether users understand why items are ordered as they are. Allocative fairness asks whether the distribution of attention and opportunity across content producers is justifiable. Collaborative agents can support fairness by assigning a dedicated critique role, but they can also obscure fairness failures behind complex deliberation. Model cards and datasheets provide documentation practices that can improve transparency, but they must be extended to cover preference profiles, agent roles, and arbitration rules [15], [16].

A further challenge is that preferences are relational and dynamic. What a user prefers today may depend on task, mood, urgency, social context, and prior exposure. A system that overfits to past behavior can trap users in filter bubbles or reduce serendipity. A system that

ignores history may fail to provide continuity. The design response is not a single optimal trade-off but an adaptive governance mechanism that allows users and institutions to inspect, adjust, and reset preference models. In public-sector settings, this may require deliberative bodies, community audits, or participatory design. In commercial settings, it may require independent audits and meaningful opt-out mechanisms. Without such mechanisms, preference-aware reranking risks becoming a sophisticated form of behavioral manipulation rather than a service to user autonomy.

6. Deployment, Infrastructure, and Sustainability

Deploying collaborative large language model agents at scale requires substantial infrastructure. Candidate generation may rely on distributed indexes, while agent deliberation may require accelerators, memory management, caching, and orchestration services. Latency budgets are tight in interactive search, so the system must decide when to invoke expensive agents and when to fall back to cheaper rankers [6], [19]. This creates a tiered architecture in which simple queries receive lightweight treatment and complex or high-stakes queries receive deeper deliberation. Such tiering can improve efficiency, but it can also create unequal service quality if some users or domains are consistently routed to lower-cost paths.

Sustainability is both an environmental and an organizational concern. Large language model inference consumes energy and water, and multi-agent deliberation multiplies inference calls. A preference-aware reranker that invokes several agents for every query may be economically and environmentally unjustifiable at scale. System designers should therefore consider caching, distillation, selective activation, and asynchronous review [13]. They should also measure total cost of ownership, including maintenance, retraining, monitoring, and incident response. The hidden technical debt of machine learning systems is well documented, and collaborative agents add new debt through prompt drift, model versioning, and emergent coordination behaviors [13], [14]. A sustainable deployment must treat these costs as first-class design constraints rather than afterthoughts.

Infrastructure choices also shape governance possibilities. Centralized deployments may simplify monitoring but concentrate power and risk. Federated or hybrid deployments may distribute control but complicate audit and consistency. Edge deployment may reduce latency but limit model capability. The appropriate architecture depends on the domain and regulatory environment. In all cases, deployment should include observability, rollback procedures, and staged rollout. The system should be able to explain why a particular ranking was produced, identify which agent versions were active, and reconstruct the preference signals that influenced the outcome. Without this infrastructure, accountability becomes aspirational rather than operational.

7. Robustness, Security, and Governance

Robustness in preference-aware reranking involves resistance to adversarial manipulation, distribution shift, and internal coordination failures. Adversarial actors may try to game rankings by injecting content, manipulating preference signals, or exploiting agent prompts. Large language models are susceptible to prompt injection, data poisoning, and jailbreak-style manipulation, which can compromise ranking integrity [12]. Multi-agent systems may be especially vulnerable because an attacker can target one role to influence the entire deliberation. Robustness therefore requires input validation, provenance checks, anomaly detection, and separation between trusted and untrusted content. It also requires red-team exercises that simulate realistic attacks on the full pipeline, not only individual models.

Governance must address both pre-deployment and post-deployment obligations. Pre-deployment governance includes impact assessments, role definition, data lineage, and approval gates. Post-deployment governance includes monitoring, incident reporting, periodic audits, and user redress. Model cards and datasheets support transparency, but collaborative reranking requires additional documentation of agent roles, arbitration logic, preference sources, and fallback behavior [15], [16]. Governance should also define who is accountable when an agent ensemble produces a harmful or discriminatory ranking. Responsibility cannot be delegated to an abstract collective of agents. Human and institutional owners must remain answerable for the system's design and operation.

A further governance challenge is cross-jurisdictional policy. Reranking systems may operate globally while being subject to conflicting laws regarding privacy, speech, competition, and discrimination. Collaborative agents may store user preferences, infer sensitive attributes, or generate rationales that reveal protected information. Policy frameworks must therefore address data minimization, purpose limitation, and cross-border transfer. They must also consider transparency rights and the ability to contest automated decisions. The technical system should be designed to support these rights through logging, explainability, and appeal workflows. Governance is not a layer added after deployment; it is a set of constraints that should shape architecture from the beginning.

8. Evaluation, Benchmarking, and Policy Implications

Evaluating preference-aware collaborative reranking is difficult because no single metric captures all relevant dimensions. Traditional retrieval metrics such as precision, recall, and normalized discounted cumulative gain remain useful for relevance, but they do not fully measure fairness, diversity, robustness, or user satisfaction [7], [20]. Evaluation should therefore combine offline benchmarks, online experiments, human studies, and adversarial testing. It should also include qualitative analysis of agent rationales and arbitration decisions. A system that ranks well on a static dataset may still fail in deployment when preferences shift, populations change, or adversarial pressure increases.

Benchmarking must also account for heterogeneity. BEIR demonstrated that models often struggle to generalize across domains, and preference-aware systems add further variation through user and institutional context [20]. New benchmarks should include multi-stakeholder preference scenarios, conflicting objectives, and governance constraints. They should record provenance, consent, and licensing for preference data. They should also test fallback behavior when agents disagree or when compute budgets are exceeded. Such benchmarks would help move the field from leaderboard optimization toward infrastructure readiness.

Policy implications follow from these evaluation gaps. Regulators may require risk assessments for high-impact ranking systems, especially when they affect access to housing, employment, credit, healthcare, or civic information. Competition policy may examine whether dominant platforms can control both preference models and the agents that interpret them. Consumer protection may address manipulative personalization and dark patterns. Public-interest technology may require open standards for preference portability and auditability. International cooperation, including the OECD AI Principles and UNESCO Recommendation on the Ethics of AI, may provide normative guidance, although implementation remains uneven. The policy challenge is to encourage beneficial personalization while preventing preference-aware reranking from becoming an unaccountable mechanism of social sorting.

9. Conclusion

Preference-aware neural reranking with collaborative large language model agents represents a significant evolution in information systems. It moves beyond static relevance scoring toward adaptive, multi-perspective ordering that can incorporate user goals, institutional constraints, and contextual nuance. This potential is real, but it depends on treating the system as a socio-technical infrastructure rather than a model-centric optimization problem. Architecture, coordination, fairness, deployment, robustness, and governance are not separate concerns; they are mutually constitutive. A poorly governed agent ensemble can amplify bias, obscure accountability, and consume unsustainable resources, while a well-governed one can support pluralistic and contestable ranking.

The paper has argued for explicit provenance, auditable coordination, meaningful user control, and lifecycle accountability. Collaborative agents should be designed with clear roles, transparent arbitration, and fallback paths. Preference models should be inspectable, resettable, and bounded by policy. Deployment should be sustainable and staged, with observability and redress. Evaluation should be heterogeneous, adversarial, and attentive to distributional impact. These commitments do not guarantee that preference-aware reranking will be beneficial, but they provide a foundation for responsible development.

Future research should examine how collaborative agent protocols behave under long-term deployment, how preference portability can be standardized, and how public institutions can participate in ranking governance. It should also investigate methods for reducing computational cost without sacrificing deliberation quality. The broader lesson is that ranking is an exercise of power over attention and opportunity. When large language model agents participate in that exercise, the design question is not only what they rank but how their judgments are legitimized. Preference-aware reranking can become a more responsive and pluralistic infrastructure, but only if it is built with governance as an architectural principle.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics.

3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates.

4. 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.

5. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). Association for Computational Linguistics.

6. Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv. https://arxiv.org/abs/1901.04085

7. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389.

8. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

9. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.

10. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

11. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv. https://arxiv.org/abs/2112.04359

12. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. https://arxiv.org/abs/1606.06565

13. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (Vol. 28, pp. 2503–2511). Curran Associates.

14. Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (pp. 1123–1132). IEEE.

15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.

16. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

17. Jin, C., Peng, H., Zhang, A., Chen, N., Zhao, J., Xie, X., ... & Metaxas, D. N. (2025, May). Rankflow: A multi-role collaborative reranking workflow utilizing large language models. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 2484-2493).

18. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., & Dou, Z. (2023). Large language models for information retrieval: A survey. arXiv. https://arxiv.org/abs/2308.07107

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

20. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (Vol. 1). Curran Associates.