

Reinforcement Learning-Based Personalized Recommendation via Collaborative LLM Reranking and User Preference Modeling

Aulikki Elo

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
contactaulikki@binghamton.edu

Yrjo Sainio

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

Abstract

Personalized recommendation has evolved from static preference estimation toward dynamic, context-sensitive decision systems that mediate attention, commerce, information access, and cultural exposure. This paper examines a system-level architecture for reinforcement learning-based personalized recommendation that combines sequential user preference modeling with collaborative large language model reranking. The central argument is that reinforcement learning offers a principled way to optimize long-horizon engagement and satisfaction, while large language models provide flexible semantic reasoning and collaborative reranking capacity that conventional recommender pipelines often lack. However, integrating these components creates substantial architectural, operational, and governance challenges. The paper analyzes retrieval, ranking, reranking, and policy layers as interdependent infrastructure rather than isolated algorithmic modules. It considers trade-offs involving latency, computational cost, explainability, fairness, robustness, privacy, and sustainability. It also discusses how collaborative reranking can improve relevance and diversity but may amplify feedback loops, popularity bias, and opaque optimization. The analysis emphasizes that deployment success depends less on any single model than on orchestration, monitoring, fallback mechanisms, human oversight, and policy-aware evaluation. Cross-domain examples from media, e-commerce, education, and public information systems illustrate how recommendation infrastructures shape user autonomy and institutional accountability. The paper concludes that reinforcement learning with collaborative language model reranking can produce adaptive and semantically rich recommendation, but only if designed as a governed socio-technical system with explicit constraints, auditable decision pathways, and continuous evaluation of fairness, robustness, and long-term societal effects.

Keywords

recommender systems, reinforcement learning, large language models, reranking, user preference modeling, fairness, governance, socio-technical systems.

1. Introduction

Personalized recommendation has become a foundational infrastructure of digital society. It determines which products consumers encounter, which news stories receive attention, which educational resources reach learners, and which cultural artifacts become visible. Early

recommender systems were largely organized around explicit rating prediction and matrix factorization, and their success established the idea that user preference could be estimated from historical interactions and used to rank items [1]. Subsequent systems incorporated deep neural networks, sequential behavior modeling, and large-scale feature engineering, enabling recommender platforms to process billions of events and update user representations continuously [2]. The field has therefore moved from static prediction toward dynamic mediation, where recommendation is not merely a matching problem but a decision-making process with temporal, economic, and ethical consequences [3].

This shift has made reinforcement learning attractive for personalized recommendation. Reinforcement learning formalizes sequential decision-making under uncertainty, allowing a system to balance immediate engagement against long-term user satisfaction, retention, and platform objectives. Unlike supervised ranking, which often optimizes a single-step label such as click or purchase, reinforcement learning can represent delayed rewards, exploration, and policy improvement over time [7], [8]. Contextual bandit methods have already been used in news recommendation and other settings where feedback is partial and counterfactual reasoning is necessary [9]. Off-policy evaluation and counterfactual risk minimization further allow historical logs to inform new policies without uncontrolled online experiments [10]. These capabilities are important because recommender systems operate in non-stationary environments where user preferences, item inventories, and social contexts change.

At the same time, large language models have introduced new possibilities for recommendation. They can represent rich semantic relationships, interpret natural-language queries, summarize user histories, and generate explanations or candidate descriptions. Retrieval-augmented generation demonstrates how language models can combine parametric knowledge with external evidence [15]. Instruction-tuned models can follow complex constraints and adapt to task formulations [13]. Reasoning-oriented prompting can support multi-step comparison and justification [14]. In recommendation, these capacities are especially relevant at the reranking stage, where a system must resolve subtle conflicts among relevance, novelty, diversity, and contextual appropriateness. A multi-role collaborative reranking workflow using large language models illustrates how different model roles can critique, revise, and aggregate candidate rankings rather than relying on a single opaque score [18]. This paper examines how such collaborative reranking can be integrated with reinforcement learning-based personalization and what system-level trade-offs follow.

The central claim is that reinforcement learning, sequential user modeling, and collaborative large language model reranking can form a powerful recommendation architecture, but only when treated as a governed socio-technical system. The architecture must coordinate retrieval, ranking, reranking, policy learning, feedback collection, monitoring, and fallback mechanisms. It must also address latency, cost, fairness, robustness, privacy, and sustainability. The remainder of the paper develops this argument by reviewing user preference modeling, reinforcement learning foundations, large language model reranking, deployment architecture, governance, sustainability, and evaluation. The goal is not to propose a narrow algorithm but to analyze the structural conditions under which adaptive recommendation can be both effective and accountable.

2. Foundations of User Preference Modeling and Sequential Recommendation

User preference modeling is the interpretive layer through which a recommender system infers what a person may want, need, or tolerate. Matrix factorization techniques demonstrated that latent factors could capture patterns in ratings and interactions, enabling

scalable personalization across sparse datasets [1]. Content-based recommendation, by contrast, uses item attributes and user profiles to match items to expressed or inferred interests, which is especially useful when interaction data are limited [17]. These approaches established two enduring ideas: preference is partially latent, and item representation matters. However, both approaches struggle with temporal dynamics, context dependence, and the difference between stated preference and actual behavior.

Deep neural recommender systems expanded the representational capacity of preference modeling. Deep neural networks for YouTube recommendation showed how candidate generation and ranking could be learned from large-scale user activity, incorporating watch history, demographic features, and contextual signals [2]. Self-attentive sequential recommendation and transformer-based models such as BERT4Rec further demonstrated that order and timing in user behavior carry substantial predictive value [4], [5]. Behavior sequence transformer models for e-commerce recommendation similarly highlighted the importance of modeling session-level intent, not only long-term taste [6]. These systems treat user preference as a trajectory rather than a fixed vector, which is essential for reinforcement learning because the policy must act on an evolving state.

Yet sequential modeling alone does not resolve the normative question of what recommendation should optimize. A system may predict the next click accurately while failing to support user well-being, diversity, or long-term trust. Preference is also socially situated. Users may click on sensational content while regretting it later, or they may rely on recommender systems to discover unfamiliar but valuable material. Content-based and knowledge-aware approaches can help by grounding recommendations in item semantics, but they may underrepresent emergent tastes and social influence [17]. Collaborative filtering and deep sequence models can capture collective behavior, but they may reproduce popularity bias and historical inequities [19]. A robust user preference model should therefore represent multiple temporal scales, including immediate intent, session goals, durable interests, and reflective preferences.

In an integrated architecture, user preference modeling supplies the state representation for reinforcement learning and the context for large language model reranking. The state may include recent interactions, long-term embeddings, declared interests, contextual constraints, and uncertainty estimates. The reranker then uses semantic reasoning to resolve ambiguous or conflicting signals. This division of labor is conceptually useful because it separates statistical personalization from interpretive judgment. However, it also creates coupling risks. If the learned state representation is biased or stale, the language model reranker may rationalize poor candidates rather than correct them. If the reranker introduces its own popularity or fluency biases, the reinforcement learning policy may misattribute reward to the wrong component. System-level design must therefore include observability across the entire pipeline, not only at the final ranking output.

3. Reinforcement Learning and Personalized Decision Making

Reinforcement learning provides a formal framework for sequential decision-making in which an agent interacts with an environment to maximize cumulative reward [8]. Its success in complex control domains demonstrated that deep reinforcement learning can learn policies from high-dimensional inputs [7]. In recommendation, the environment includes users, items, platform interfaces, and competing services. The reward may include clicks, purchases, watch time, satisfaction surveys, retention, or explicit feedback. The challenge is that these signals are noisy, delayed, and often only partially observed. A click does not necessarily imply

satisfaction, and a purchase may occur after multiple sessions and external influences. Reinforcement learning can model such delayed effects, but only if the reward function and environment model are carefully designed.

Contextual bandits offer a middle ground between supervised learning and full reinforcement learning. They model each recommendation as a decision with observable context and partial feedback, making them suitable for news recommendation and other settings where long-term transition modeling is difficult [9]. Off-policy learning allows systems to evaluate new policies using logged data, which is important because online experimentation can be costly or harmful [10]. However, off-policy evaluation depends on assumptions about how data were collected. If the logging policy was highly selective, counterfactual estimates may be biased. This matters in personalized recommendation because the system itself shapes the data it later learns from, creating feedback loops.

The exploration-exploitation trade-off is central. A system that exploits known preferences may improve short-term engagement but reduce discovery and entrench existing tastes. A system that explores too aggressively may frustrate users or expose them to low-quality content. In high-stakes domains such as education, health, or public information, exploration must be constrained by safety and appropriateness. Reward shaping is similarly consequential. If the reward emphasizes click-through rate, the system may favor clickbait. If it emphasizes watch time, it may favor compulsive consumption. If it emphasizes diversity, it may reduce immediate satisfaction unless users are given control and transparency. Reinforcement learning does not solve these normative choices; it makes them explicit in the objective function and policy update.

From a system perspective, reinforcement learning introduces infrastructure requirements. The system needs reliable logging, delayed reward attribution, simulation or replay environments, policy versioning, and safety constraints. It also needs mechanisms to roll back harmful policies and to detect distribution shift. In an architecture with large language model reranking, the action space is not limited to item selection. The system may also choose explanations, presentation formats, or conversational prompts. This expanded action space increases flexibility but also complicates credit assignment. If a user engages more after a language model generates a persuasive explanation, it is unclear whether the reward should update the retrieval model, the ranking policy, the reranker, or the explanation generator. Governance must therefore include component-level accountability and causal analysis.

4. Large Language Models and Collaborative Reranking

Large language models have changed the design space of recommendation by providing general-purpose semantic reasoning. Foundation models can transfer across tasks and domains, enabling zero-shot and few-shot adaptation [12]. Instruction tuning aligns model behavior with user intent and task instructions [13]. Chain-of-thought prompting improves performance on tasks requiring multi-step reasoning [14]. Retrieval-augmented generation combines model knowledge with external documents, which can improve factuality and updateability [15]. Neural ranking models for document retrieval also demonstrate how deep semantic matching can improve relevance beyond lexical overlap [16]. Together, these developments make language models attractive for the reranking stage, where candidates must be compared according to nuanced criteria.

Collaborative reranking extends this idea by using multiple model roles or perspectives. Instead of asking one model to produce a final list, the system may assign roles such as

relevance assessor, diversity advocate, safety reviewer, and user-context interpreter. A multi-role collaborative reranking workflow using large language models shows how critique and aggregation can improve ranking decisions by making explicit the trade-offs that a single score may hide [18]. This approach is valuable because recommendation is inherently multi-objective. Relevance, novelty, serendipity, timeliness, and appropriateness often conflict. A collaborative workflow can produce a more reasoned ranking, but it also increases computational cost and complexity.

The integration of collaborative reranking with reinforcement learning creates a layered policy architecture. The base recommender retrieves and ranks candidates using efficient models. The large language model reranker then reorders or filters candidates based on semantic and contextual reasoning. The reinforcement learning policy may operate above this stack, selecting among reranking strategies, adjusting exploration, or personalizing the weighting of objectives. Alternatively, reinforcement learning may fine-tune the reranker itself using user feedback. Both designs have trade-offs. Policy-over-strategies is easier to monitor and roll back, while policy-inside-reranker may achieve finer adaptation but is harder to audit.

Large language model reranking also introduces new failure modes. Models may hallucinate item attributes, favor fluent but irrelevant candidates, or exhibit position bias. They may be sensitive to prompt phrasing and may produce inconsistent outputs across runs. Collaborative roles can reduce some errors but may also amplify agreement bias if roles are not genuinely independent. Latency is another concern because multi-role reranking may require multiple model calls. Caching, distillation, and routing can mitigate cost, but they may reduce adaptivity. The system must therefore treat the reranker as a probabilistic component with uncertainty, not as an oracle. Fallback to conventional ranking, confidence thresholds, and human review for high-stakes cases are essential safeguards.

5. System Architecture and Deployment Trade-offs

A production recommendation architecture that combines reinforcement learning and collaborative large language model reranking must coordinate several layers. The retrieval layer selects a manageable candidate set from a large corpus using embeddings, inverted indexes, or graph-based methods. The ranking layer assigns preliminary scores using deep models trained on interaction data [2], [6]. The reranking layer applies semantic reasoning and collaborative evaluation [18]. The policy layer determines exploration, objective weighting, and long-term optimization. A feedback layer logs outcomes, delays, and contextual signals. An observability layer monitors drift, fairness, latency, and failure rates. Each layer has its own model lifecycle, data dependencies, and failure modes.

Deployment trade-offs are unavoidable. Multi-role language model reranking can improve quality but increases latency and cost. If the system has strict response-time requirements, it may need to rerank only the top few candidates or use smaller distilled models for most traffic and larger models for uncertain cases. Reinforcement learning may improve long-term outcomes but requires exploration, which can degrade short-term user experience. A conservative deployment might use offline policy evaluation [10] and shadow deployment before live experimentation. It might also use contextual bandits for limited exploration [9] and reserve full reinforcement learning for domains where delayed rewards are well understood.

Infrastructure also matters for robustness. Language model inference can fail due to rate limits, version changes, or network issues. The recommender must have graceful degradation strategies. If the reranker is unavailable, the system should fall back to a deterministic ranking. If the reinforcement learning policy is uncertain, it should default to safe, well-understood behavior. If feedback signals are delayed, the system should avoid premature updates. These are not merely engineering details; they determine whether the system remains stable under real-world conditions. A recommendation architecture that cannot degrade gracefully may produce outages, harmful recommendations, or loss of user trust.

Organizational factors shape architecture as well. Teams responsible for retrieval, ranking, language models, and policy may have different objectives and incentives. A collaborative reranking workflow may cross team boundaries, requiring shared evaluation metrics and incident response procedures. Governance must define who can change the reward function, who approves new model roles, and who audits fairness outcomes. Without such coordination, the system may optimize locally while degrading globally. The most sophisticated model stack can fail if operational ownership is fragmented.

6. Robustness, Fairness, Governance, and Policy Implications

Recommender systems are not neutral mirrors of preference. They allocate visibility, shape demand, and influence beliefs. Large language models introduce additional risks because they can generate persuasive text and may reproduce biases present in training data [11]. Foundation models also concentrate capability and infrastructure, raising questions about accountability, access, and market power [12]. When such models are used for reranking, their judgments may appear objective while encoding contested values. Governance must therefore address both algorithmic fairness and institutional legitimacy.

Fairness in recommendation includes distributional concerns such as exposure for new or minority items, equal treatment across user groups, and avoidance of harmful stereotypes. Popularity bias is a well-documented issue, where popular items receive disproportionate visibility and feedback loops reinforce their advantage [19]. Reinforcement learning can worsen this problem if reward is dominated by aggregate engagement. Collaborative reranking can worsen it if language models favor well-known or well-described items. Conversely, these techniques can also mitigate bias if diversity, calibration, and representation are explicitly included in reranking criteria and reward functions. The key is that fairness must be designed into the objective and monitored continuously, not treated as an afterthought.

Governance requires transparency and auditability. Users should have meaningful explanations of why items are recommended, especially when language models generate justifications. Regulators and auditors need access to model cards, evaluation reports, and change logs. Privacy must be protected through data minimization, purpose limitation, and secure computation. Consent is complicated in recommendation because preferences are inferred from behavior, and users may not understand the extent of profiling. Policy frameworks such as fairness through awareness provide conceptual foundations for treating individuals according to relevant attributes while avoiding unjustified discrimination [20]. In practice, governance must combine technical audits, external oversight, and user controls.

High-stakes domains require stricter constraints. In education, a recommendation system should not optimize engagement at the expense of learning outcomes. In health, it must avoid harmful misinformation and respect clinical boundaries. In public information, it should

support informed citizenship rather than polarization. Reinforcement learning can optimize long-term objectives in these domains, but only if the reward function reflects domain values and if exploration is bounded. Collaborative language model reranking can incorporate policy documents, ethical guidelines, and expert review, but it cannot replace institutional accountability. The system should be designed to escalate uncertain or sensitive cases to human decision-makers.

7. Sustainability and Operational Economics

The computational cost of large language models and reinforcement learning is substantial. Training foundation models consumes significant energy and resources [11], and serving multi-role rerankers at scale can increase operational emissions. Reinforcement learning adds further cost through simulation, experimentation, and frequent policy updates. Sustainability is therefore not peripheral but constitutive of responsible recommendation infrastructure. A system that improves relevance marginally while greatly increasing energy use may be socially inefficient. Operators should evaluate the marginal benefit of each component against its carbon and financial cost.

Several architectural choices can improve sustainability. Distillation can compress large rerankers into smaller models for common cases. Caching can reuse reranking decisions for similar contexts, though it risks staleness. Routing can send only uncertain or high-value requests to multi-role language models. Retrieval-augmented generation can reduce the need for retraining by updating external knowledge [15]. Reinforcement learning can prioritize sample-efficient methods and offline evaluation to reduce wasteful online experimentation [10]. These choices must be balanced against fairness and robustness. A cheap system that systematically harms minority users is not sustainable in a normative sense.

Operational economics also shape access. If collaborative large language model reranking is only affordable to large platforms, smaller services and public institutions may be disadvantaged. This could concentrate recommendation power and reduce pluralism. Open models, shared evaluation benchmarks, and public infrastructure can mitigate this risk. Policy interventions might include transparency requirements for algorithmic curation, support for public-interest recommendation, and funding for independent auditing. The goal is not to eliminate personalization but to ensure that its benefits and costs are distributed fairly.

8. Evaluation, Measurement, and Cross-Domain Transfer

Evaluating reinforcement learning-based recommendation with collaborative language model reranking is difficult because the system optimizes long-term outcomes in a non-stationary environment. Offline metrics such as precision, recall, and normalized discounted cumulative gain are insufficient. They cannot capture delayed satisfaction, diversity, or strategic user behavior. Counterfactual evaluation can estimate policy performance from logged data [10], but it depends on assumptions that may not hold. Online experiments provide stronger evidence, but they raise ethical concerns, especially when exploration may expose users to harmful content. A mixed evaluation strategy should combine offline diagnostics, simulation, shadow deployment, limited online trials, and long-term user studies.

Cross-domain transfer is another challenge. A model trained on e-commerce may not transfer well to education or news because the reward structures and user goals differ. Large language models offer some zero-shot generalization [12], and instruction tuning can adapt behavior to new tasks [13]. Retrieval-augmented generation can incorporate domain-specific knowledge [15]. Neural ranking models can improve semantic matching across domains [16]. However,

transfer does not eliminate the need for domain-specific validation. Fairness criteria, privacy rules, and safety thresholds vary by context. A system that performs well in entertainment may be inappropriate in health or public information.

Evaluation should also examine component interactions. A reinforcement learning policy may improve engagement while the language model reranker increases perceived relevance but reduces diversity. A conventional ranking model may be more stable than a multi-role reranker but less able to handle complex queries. The system should track end-to-end outcomes and component-level contributions. This requires causal inference methods, ablation studies, and detailed logging. It also requires organizational commitment to continuous monitoring. Recommendation quality is not a static property; it evolves with users, items, and society.

9. Conclusion

This paper has examined reinforcement learning-based personalized recommendation that integrates collaborative large language model reranking and user preference modeling. The analysis suggests that these components can complement one another. Sequential and deep preference models provide rich state representations [1], [2], [4], [5], [6]. Reinforcement learning provides a framework for long-term decision-making under uncertainty [7], [8], [9], [10]. Large language models provide semantic reasoning and flexible reranking capacity [12], [13], [14], [15], [16]. Collaborative reranking can make multi-objective trade-offs more explicit [18]. Together, they enable adaptive, context-aware recommendation that goes beyond static ranking.

However, integration creates serious system-level challenges. Latency, cost, feedback loops, bias, opacity, and sustainability must be addressed through architecture and governance, not merely through model improvement. Fairness requires explicit objectives and continuous auditing [19], [20]. Robustness requires fallback mechanisms, uncertainty handling, and graceful degradation. Sustainability requires efficient routing, distillation, caching, and careful evaluation of marginal benefits. Governance requires transparency, privacy protection, human oversight, and policy frameworks that hold platforms accountable. The most important design principle is that recommendation infrastructure should be treated as a socio-technical system whose outputs shape user autonomy and social outcomes.

Future work should develop standardized evaluation suites for reinforcement learning with language model reranking, including measures of long-term satisfaction, fairness, robustness, and environmental cost. It should also explore participatory design methods that give users and affected communities a role in defining recommendation objectives. Cross-domain studies are needed to understand when collaborative reranking improves outcomes and when it introduces new risks. The promise of reinforcement learning-based personalized recommendation is significant, but it can only be realized if technical innovation is matched by institutional responsibility and sustained public scrutiny.

References

1. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37. <https://doi.org/10.1109/MC.2009.263>
2. Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender*

Systems (pp. 191-198). Association for Computing Machinery.
<https://doi.org/10.1145/2959100.2959190>

3. Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender systems: Techniques, applications, and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (3rd ed., pp. 1-35). Springer. https://doi.org/10.1007/978-1-0716-2197-4_1
4. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441-1450). Association for Computing Machinery. <https://doi.org/10.1145/3357384.3357895>
5. Zhou, D., & Yi, X. (2026). DISTRIBUTED AI-ASSISTED RISK ASSESSMENT FOR ZERO-TRUST MICROSERVICES: A TAXONOMY AND CONCEPTUAL FRAMEWORK. *AI and Data Science Journal*, 3(2), 1-13.
6. Chen, Q., Zhao, H., Li, W., Huang, P., & Ou, W. (2019). Behavior sequence transformer for e-commerce recommendation in Alibaba. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data* (pp. 1-4). Association for Computing Machinery. <https://doi.org/10.1145/3326937.3341261>
7. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533. <https://doi.org/10.1038/nature14236>
8. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
9. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 661-670). Association for Computing Machinery. <https://doi.org/10.1145/1772690.1772758>
10. Swaminathan, A., & Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 814-823). PMLR.
11. 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). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
12. 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>
13. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens,

- M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
14. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.
 15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
 16. Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 55-64). Association for Computing Machinery. <https://doi.org/10.1145/2983323.2983769>
 17. Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (pp. 325-341). Springer. https://doi.org/10.1007/978-3-540-72079-9_10
 18. Jin, Can, et al. "Rankflow: A multi-role collaborative reranking workflow utilizing large language models." *Companion Proceedings of the ACM on Web Conference 2025*. 2025.
 19. Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems* (pp. 42-46). Association for Computing Machinery. <https://doi.org/10.1145/3109859.3109912>
 20. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). Association for Computing Machinery. <https://doi.org/10.1145/2090236.2090255>