

Personalized Retrieval-Augmented Generation for User-Aware Conversational Recommendation Systems

Cianan R. Carrigan

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

ciananc@unr.edu

Ferdinando Gemitto

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

ferdinandowork@buffalo.edu

Kajsa Holmstrom

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

holmstromkajsa@ku.edu

Mart Massa

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.

martm@uc.edu

Abstract

Personalized retrieval-augmented generation has emerged as a promising paradigm for conversational recommendation systems, yet its system-level implications remain underexamined. This paper presents a critical and architectural analysis of user-aware retrieval-augmented generation, focusing on how personalization layers interact with retrieval infrastructure, ranking mechanisms, governance frameworks, and long-term sustainability. The central argument is that effective personalization in conversational recommendation is not merely a modeling problem but a socio-technical design challenge that requires careful balancing of relevance, fairness, robustness, latency, and privacy. The paper first reviews foundations of retrieval-augmented generation and conversational recommendation, then develops a layered architecture that separates global retrieval, user-conditioned re-ranking, and dialogue-state fusion. It examines indexing and deployment trade-offs, including approximate nearest neighbor search, sharding strategies, and hardware acceleration, and discusses how these choices affect responsiveness and operational cost. The analysis then turns to governance, showing how user-aware retrieval can amplify feedback loops, filter bubbles, and privacy risks, and how policy mechanisms such as data minimization, auditability, and contestability must be embedded into system design. Robustness and evaluation are addressed through the lens of distribution shift, adversarial manipulation, and multi-objective assessment. Case illustrations from e-commerce, media streaming, and public service navigation demonstrate cross-domain variation in requirements and constraints. The paper concludes with forward-looking perspectives on federated personalization, adaptive retrieval control, and sustainable infrastructure, arguing that future progress depends on

integrating technical, organizational, and policy considerations into a coherent system-level agenda.

Keywords

personalized retrieval, retrieval-augmented generation, conversational recommendation, user-aware systems, socio-technical infrastructure, fairness and governance, approximate nearest neighbor search, sustainable deployment.

1. Introduction

Retrieval-augmented generation has rapidly moved from a niche technique for knowledge-intensive natural language processing to a general architectural pattern for grounding generative models in external evidence [1]. Early formulations demonstrated that coupling a parametric language model with a non-parametric memory of documents could improve factuality and domain adaptability without retraining the entire model [2]. Subsequent work refined dense retrieval, fusion-in-decoder architectures, and large-scale retrieval from trillion-token corpora, establishing retrieval as a first-class component of modern generative systems [3], [4], [5]. Surveys of the field now describe a rich design space spanning retrieval timing, query formulation, evidence selection, and generation conditioning [6]. Within this space, conversational recommendation represents a particularly demanding application because the system must not only retrieve relevant items or supporting knowledge but also maintain a coherent dialogue, infer user preferences from sparse signals, and adapt recommendations over multiple turns.

Conversational recommendation systems have traditionally been studied through the lens of interactive preference elicitation, sequential modeling, and dialogue management [13], [14], [15]. Neural approaches have improved the ability to model user-item interactions, temporal dynamics, and session-level intent [16], [17]. However, the integration of retrieval-augmented generation into these systems introduces new system-level tensions. Retrieval must be personalized without collapsing into overfitting, generation must remain grounded without becoming repetitive, and the overall pipeline must operate within strict latency and cost budgets. Moreover, because recommendations influence what users see, buy, read, and believe, the governance of personalized retrieval is inseparable from questions of fairness, privacy, and accountability [21], [22]. This paper argues that user-aware retrieval-augmented generation for conversational recommendation should be understood as a layered socio-technical infrastructure rather than a single model.

The remainder of the paper is organized as follows. Section 2 reviews foundations of retrieval-augmented generation. Section 3 frames conversational recommendation as a socio-technical system. Section 4 develops a layered architecture for user-aware personalized retrieval. Section 5 examines indexing, infrastructure, and latency trade-offs. Section 6 addresses governance, fairness, privacy, and policy. Section 7 discusses robustness, evaluation, and sustainability. Section 8 provides case illustrations and cross-domain comparisons. Section 9 outlines future directions, and Section 10 concludes.

2. Foundations of Retrieval-Augmented Generation

The core idea of retrieval-augmented generation is to separate knowledge storage from knowledge manipulation. A retriever selects relevant evidence from an external corpus, and a generator conditions on that evidence to produce an output [1]. This separation offers several advantages: the corpus can be updated without retraining the generator, provenance can be

traced to specific documents, and the model can access long-tail information that would be impractical to memorize. Early systems demonstrated these benefits in open-domain question answering [2], [3]. Dense passage retrieval showed that learned embeddings could outperform sparse lexical matching for many semantic retrieval tasks [3]. Fusion-in-decoder architectures further improved generation by allowing the model to attend jointly to multiple retrieved passages [4]. Retrieval from very large corpora extended the approach to trillions of tokens, suggesting that scale alone can yield substantial gains when combined with efficient indexing [5].

As the field matured, attention shifted from whether retrieval helps to when and how retrieval should be triggered. Some systems retrieve once per query, while others retrieve iteratively or adaptively [7], [8]. Self-reflective approaches train models to critique their own generations and decide whether additional retrieval is needed [7]. Active retrieval methods allow the model to formulate follow-up queries when initial evidence is insufficient [8]. Black-box retrieval augmentation enables off-the-shelf language models to benefit from retrieved evidence without fine-tuning [9]. These developments reflect a broader recognition that retrieval is not a static preprocessing step but a dynamic control problem. In conversational settings, this control problem becomes even more acute because the system must decide what to retrieve, when to retrieve it, and how to present it within an ongoing dialogue.

A parallel line of work has focused on the representational and infrastructural substrate of retrieval. Sentence-level embeddings provide a compact semantic representation that can be indexed and searched efficiently [10]. Hierarchical navigable small world graphs enable approximate nearest neighbor search at scale with strong recall-latency trade-offs [11]. GPU-accelerated libraries have made billion-scale similarity search practical for production systems [12]. These infrastructural advances are not merely implementation details; they shape what kinds of personalization are feasible. For example, if retrieval latency is dominated by index traversal, then user-conditioned re-ranking must be extremely lightweight. If embedding updates are expensive, then personalization may need to be expressed through query transformation rather than index modification. Understanding these constraints is essential for designing user-aware retrieval-augmented generation.

3. Conversational Recommendation as a Socio-Technical System

Conversational recommendation differs from single-shot recommendation in that the system must engage in a multi-turn exchange to uncover user preferences, explain suggestions, and adapt to feedback [20]. Classical recommender systems often rely on matrix factorization and neural collaborative filtering to model user-item interactions [13], [14]. These approaches excel at capturing latent preference structures but typically assume a static interaction context. Sequential recommendation models such as self-attentive and bidirectional transformer architectures improve the ability to model temporal order and session dynamics [15], [16]. Behavior sequence transformers further incorporate contextual features such as time, location, and device [17]. Together, these models provide a strong foundation for understanding user behavior, but they do not inherently solve the retrieval and grounding problems that arise when a generative model is placed at the center of the interaction.

A socio-technical perspective highlights that conversational recommendation systems are embedded in organizational workflows, business models, and regulatory environments. The choice of what to recommend is never purely technical; it reflects commercial objectives, editorial policies, and legal constraints. In e-commerce, for example, recommendations may be optimized for conversion, margin, or long-term customer satisfaction. In media streaming,

the objective may be engagement, retention, or diversity of exposure. In public service navigation, the objective may be accuracy, accessibility, and equitable access to information. These differing objectives imply that the same retrieval-augmented generation architecture may need different governance mechanisms, evaluation metrics, and fail-safes. A system that performs well on offline relevance metrics may still fail in practice if it cannot explain its recommendations, if it exposes sensitive inferences, or if it systematically disadvantages certain user groups.

The integration of retrieval-augmented generation into conversational recommendation also changes the locus of personalization. In traditional systems, personalization is often encoded in user and item embeddings that are learned during training. In retrieval-augmented systems, personalization can be expressed through the retrieval query, the ranking function, the evidence selection policy, or the generation prompt. This flexibility is powerful but also introduces new failure modes. A personalized retrieval query may overfit to a user's recent behavior and miss serendipitous items. A personalized ranker may amplify popular or polarizing content. A personalized generation prompt may leak private information into the dialogue. Addressing these risks requires architectural choices that make personalization explicit, auditable, and controllable rather than implicit and opaque.

4. Architecture of User-Aware Personalized Retrieval

A practical architecture for user-aware retrieval-augmented conversational recommendation can be organized into three layers: global retrieval, user-conditioned re-ranking, and dialogue-state fusion. The global retrieval layer maintains a shared index of items, documents, or knowledge passages [10], [11], [12]. This layer is responsible for broad recall and is typically optimized for throughput and scalability. It may use dense embeddings, sparse lexical representations, or hybrid approaches. The user-conditioned re-ranking layer takes the candidate set from global retrieval and applies a personalized scoring function. This layer is where user history, session context, and inferred preferences are brought to bear. The dialogue-state fusion layer integrates the ranked candidates with the conversational context, generating responses that are grounded, coherent, and appropriately personalized. Separating these layers makes it possible to reason about trade-offs independently and to assign clear ownership for different aspects of system behavior.

The user-conditioned re-ranking layer is the most direct locus of personalization in this architecture. It can incorporate collaborative filtering signals, sequential behavior embeddings, and contextual features [13], [14], [15], [16], [17]. However, it must also contend with the cold-start problem, noise in user feedback, and the risk of feedback loops. A user who frequently engages with a narrow category may be recommended more of the same, reducing exposure to diverse content. A user who clicks on a misleading item may inadvertently signal a preference that distorts future recommendations. To mitigate these risks, re-ranking should be designed with explicit diversity, novelty, and fairness objectives, and it should be able to fall back to non-personalized ranking when confidence is low. Recent work on unified semantic spaces and attention-based fusion across heterogeneous sources offers a useful structural template for reconciling these tensions, although its focus remains on retrieval quality rather than user-aware ranking [18]. Extending such frameworks to incorporate user-conditioned signals and governance constraints is a promising direction.

Dialogue-state fusion is where the system translates retrieved and ranked evidence into a conversational response. This layer must decide how much retrieved evidence to include, how to phrase recommendations, and how to handle ambiguity or disagreement. It must also

manage the trade-off between fluency and faithfulness: a response that is too fluent may hallucinate details, while a response that is too literal may be repetitive and unhelpful. Techniques such as self-reflection and active retrieval can help, but they add latency and computational cost [7], [8]. In a conversational setting, the system may also need to ask clarifying questions, acknowledge uncertainty, or explain why a particular item was recommended. These behaviors require the dialogue-state fusion layer to have access to provenance information, confidence estimates, and policy constraints. Without such information, the system may generate plausible but ungrounded recommendations, undermining user trust and organizational accountability.

5. Indexing, Infrastructure, and Latency Trade-offs

The performance of user-aware retrieval-augmented generation depends heavily on the underlying indexing and serving infrastructure. Approximate nearest neighbor search methods such as hierarchical navigable small world graphs provide sublinear query times with tunable recall [11]. GPU-accelerated similarity search enables high-throughput retrieval over billions of vectors [12]. These techniques make it possible to retrieve relevant candidates in milliseconds, but they also introduce trade-offs. Higher recall often requires more memory and longer query times. Larger indexes may need to be sharded across multiple machines, introducing network latency and coordination overhead. Embedding updates may require partial index rebuilds, which can be expensive and disruptive. In a conversational system, these trade-offs directly affect user experience: a delay of a few hundred milliseconds can make a dialogue feel sluggish, while a recall failure can produce an irrelevant or repetitive response.

Deployment choices further shape the design space. Centralized retrieval offers consistency and easier governance but may struggle with privacy requirements and geographic latency. Federated or edge-based retrieval can reduce latency and keep sensitive data local, but it complicates index maintenance and global personalization. Caching strategies can improve responsiveness for frequent queries, but they may also reinforce popularity bias and reduce exposure to long-tail items. Sharding by user segment can improve personalization but may create isolated indexes that are difficult to update and audit. These are not merely engineering details; they are architectural decisions with governance implications. For example, if user data is distributed across edge nodes, then auditing for fairness or compliance becomes more complex. If caching is aggressive, then explanations for why a particular recommendation appeared may be difficult to reconstruct.

Cost is another critical dimension. Retrieval-augmented generation can be computationally expensive because it involves embedding computation, index traversal, re-ranking, and generation. In high-traffic conversational systems, these costs accumulate quickly. Surveys of neural recommendation and conversational recommendation provide useful taxonomies of methods and their computational characteristics [19], [20]. However, they also highlight that many systems are evaluated in offline settings that do not reflect production constraints. A system that achieves state-of-the-art accuracy in a research benchmark may be economically infeasible at scale. Conversely, a system that is efficient may sacrifice personalization quality or fairness. Negotiating these trade-offs requires a clear understanding of the service-level objectives, the cost structure, and the acceptable latency envelope. It also requires ongoing monitoring and adaptation, because user behavior, content catalogs, and hardware availability change over time.

6. Governance, Fairness, Privacy, and Policy

User-aware retrieval-augmented generation raises substantial governance challenges. Because recommendations can shape attention, consumption, and belief, they are subject to ethical and regulatory scrutiny. Language models trained on large corpora can encode and amplify biases, and retrieval can surface biased or harmful content if the index is not carefully curated [21], [22]. Personalization can compound these risks by creating filter bubbles, reinforcing stereotypes, and limiting exposure to diverse perspectives. In conversational recommendation, the risks are amplified because the system may infer sensitive attributes from dialogue and use them to tailor recommendations without explicit user consent. Governance must therefore address not only the content of the index but also the logic of personalization and the transparency of the interaction.

Privacy is a central concern. User-aware retrieval often relies on detailed behavioral histories, session context, and inferred preferences. These data can be sensitive, and their retention and use may be subject to legal restrictions such as data protection regulations. Data minimization, purpose limitation, and user control are established principles that must be translated into system design. For example, the system might store only aggregated preference signals rather than raw interaction logs, or it might allow users to inspect and delete their personalization profiles. Differential privacy and federated learning can provide technical safeguards, but they also impose costs in terms of accuracy and complexity. A pragmatic approach is to treat privacy as a design constraint from the outset, rather than as a compliance checkbox added after deployment.

Auditability and contestability are equally important. When a conversational system recommends a product, article, or service, users and regulators may want to know why. This requires provenance tracking, explanation generation, and the ability to challenge or override recommendations. In a retrieval-augmented system, provenance can be traced to specific retrieved documents or items, which is an advantage over purely parametric models. However, the ranking and fusion layers may obscure the relationship between the retrieved evidence and the final recommendation. Making these layers interpretable, or at least auditable, is a significant engineering challenge. It also requires organizational processes for reviewing and responding to complaints. Without such processes, transparency mechanisms may be performative rather than substantive.

7. Robustness, Evaluation, and Sustainability

Robustness is a critical requirement for conversational recommendation systems. Users may behave unpredictably, adversarial actors may attempt to manipulate retrieval or ranking, and the underlying content distribution may shift over time. A system that performs well on historical data may fail when confronted with new items, new user segments, or coordinated manipulation. Robustness can be improved through techniques such as adversarial training, outlier detection, and conservative fallback policies. However, these techniques often trade off against personalization quality. A system that is too conservative may fail to adapt to legitimate changes in user preference; a system that is too aggressive may be easily gamed. Striking the right balance requires continuous monitoring and human oversight, as well as the ability to roll back changes quickly.

Evaluation must also be multi-objective. Accuracy and relevance are important, but they are not sufficient. Conversational recommendation systems should be evaluated on diversity, novelty, serendipity, fairness, latency, and user satisfaction. Offline metrics such as recall and normalized discounted cumulative gain provide useful signals, but they may not correlate with online outcomes [19], [20]. Online experiments, user studies, and long-term retention

analyses are necessary to understand the real-world impact of design choices. Moreover, evaluation should be conducted across different user groups to detect disparate performance. A system that improves average accuracy but harms a minority group may be unacceptable from a governance perspective. Multi-objective evaluation frameworks can help surface these trade-offs, but they require organizational commitment to acting on the results.

Sustainability is an often-overlooked dimension of retrieval-augmented generation. Training and serving large models consume significant energy, and the environmental impact can be substantial [23], [24]. Retrieval adds additional computational overhead, particularly if the index is large and queries are frequent. Sustainable design involves optimizing model size, retrieval frequency, and hardware utilization, as well as considering the lifecycle costs of data storage and index maintenance. It also involves questioning whether personalization always requires the most expensive techniques. In some cases, simpler heuristics or cached responses may be sufficient. Sustainability should be treated as a first-class design objective, alongside accuracy and fairness. This means measuring energy consumption, reporting it transparently, and making architectural choices that reduce waste without compromising user value.

8. Case Illustrations and Cross-Domain Comparisons

Consider an e-commerce conversational assistant that helps users find products. The system must retrieve from a large catalog, personalize based on purchase history and browsing behavior, and generate persuasive but accurate descriptions. Latency is critical because users expect near-instant responses. Fairness concerns include whether certain sellers or product categories are systematically disadvantaged. Governance requires transparency about sponsored recommendations and compliance with consumer protection rules. A media streaming assistant, by contrast, may prioritize engagement and retention, but it must also manage diversity of exposure and avoid reinforcing filter bubbles. Retrieval may involve both content metadata and user-generated reviews. The dialogue may be more open-ended, and the system may need to handle sensitive topics such as news or political content. These differences illustrate that the same architectural pattern must be parameterized differently across domains.

Public service navigation provides a third illustration. A conversational system that helps citizens find benefits, healthcare, or legal aid must prioritize accuracy, accessibility, and equity. Personalization may be less about commercial preference and more about language, location, and eligibility. Retrieval must be grounded in authoritative documents, and errors can have serious consequences. Governance requirements are stringent, including auditability, privacy, and non-discrimination. In such settings, the system may need to defer to human agents when uncertainty is high. These case comparisons show that user-aware retrieval-augmented generation is not a one-size-fits-all solution. The architecture must be adaptable, and the governance framework must be tailored to the domain. Cross-domain learning can still be valuable, but it must be accompanied by careful contextualization.

9. Future Directions

Several promising directions lie ahead. Federated personalization could allow user-aware retrieval without centralizing sensitive data. Adaptive retrieval control could enable systems to decide dynamically when personalization is necessary and when a generic response is sufficient. Multi-objective optimization could help balance relevance, diversity, fairness, and latency in a principled way. Explainability and contestability could be improved through provenance-aware generation and user-facing controls. Sustainability could be enhanced

through model compression, efficient indexing, and energy-aware scheduling. Each of these directions involves trade-offs, and none is a silver bullet. Progress will depend on interdisciplinary collaboration among computer scientists, engineers, social scientists, legal scholars, and policymakers.

Another important direction is the development of shared benchmarks and evaluation protocols that reflect production constraints. Many current benchmarks focus on offline accuracy and do not capture latency, cost, fairness, or robustness. Building realistic testbeds that simulate user behavior, adversarial attacks, and governance requirements would help researchers and practitioners make more informed decisions. Such testbeds should also include mechanisms for auditing and red-teaming. Finally, the field needs to grapple with the long-term societal implications of personalized conversational recommendation. As these systems become more capable and more pervasive, their influence on human attention, choice, and autonomy will grow. Responsible design requires not only technical excellence but also ongoing public deliberation and adaptive regulation.

10. Conclusion

Personalized retrieval-augmented generation for user-aware conversational recommendation systems is a complex, multi-layered enterprise. It combines advances in retrieval, generation, recommendation, and dialogue management with infrastructural, governance, and sustainability considerations. This paper has argued that effective personalization cannot be reduced to a modeling problem. It requires an architectural separation of concerns, careful attention to latency and cost, robust evaluation across multiple objectives, and governance mechanisms that address fairness, privacy, and accountability. The proposed layered architecture, comprising global retrieval, user-conditioned re-ranking, and dialogue-state fusion, provides a useful framework for reasoning about these issues. It also highlights the need for ongoing monitoring, human oversight, and the ability to adapt to changing conditions.

Looking forward, the field should prioritize interdisciplinary research that connects technical design with policy and organizational practice. Federated and adaptive approaches, multi-objective optimization, and sustainable infrastructure are promising avenues, but they must be developed with a clear understanding of their trade-offs. Ultimately, the success of personalized retrieval-augmented conversational recommendation will be measured not only by its accuracy or efficiency but by its ability to serve users fairly, transparently, and sustainably. That is a system-level challenge, and it demands a system-level response.

References

1. 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.
2. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 3929-3938). PMLR.
3. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (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.

4. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 874-880). Association for Computational Linguistics.
5. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J. B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (pp. 2206-2240). PMLR.
6. 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 preprint arXiv:2312.10997.
7. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the 12th International Conference on Learning Representations. OpenReview.
8. Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 7969-7992). Association for Computational Linguistics.
9. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Liang, P., & Yih, W. T. (2023). REPLUG: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.
10. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3982-3992). Association for Computational Linguistics.
11. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824-836.
12. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547.
13. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.
14. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (pp. 173-182). Association for Computing Machinery.
15. Kang, W. C., & McAuley, J. (2018). Self-attentive sequential recommendation. In Proceedings of the 2018 IEEE International Conference on Data Mining (pp. 197-206). IEEE.
16. 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.

17. 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.
18. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In 2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 591-597). IEEE.
19. Wu, L., He, X., Wang, X., Zhang, K., & Wang, M. (2022). A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5), 4425-4445.
20. Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5), 1-36.
21. 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.
22. 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 preprint arXiv:2112.04359*.
23. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics.
24. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54-63.