

# Memory-Augmented Large Language Models for Adaptive E-Commerce Recommendation Systems

Kenneth D. Woods

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

woods210@oregonstate.edu

Brant Herrisan

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

brent.harrison@ucf.edu

Matteo Simmons

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

hellomatteo@unh.edu

Mikkel M. Ferguson

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

mferguson@uc.edu

## Abstract

Large language models have introduced new opportunities for e-commerce recommendation systems by enabling semantic reasoning over user behavior, product descriptions, and interaction contexts. However, the parametric memory of static models is poorly suited to the rapidly shifting preferences and session-specific constraints that characterize real-world retail platforms. Memory-augmented large language models address this limitation by combining parametric generation with persistent user-specific memory stores, retrieval mechanisms, and adaptive update policies. This paper provides a system-level examination of memory-augmented large language models for adaptive e-commerce recommendation. It analyzes architectural foundations, memory lifecycle management, structural trade-offs between retrieval and generation, and the implications for latency, scalability, and correctness. The discussion further addresses robustness, fairness, governance, sustainability, and deployment constraints. Rather than focusing on narrow model-level metrics, the paper foregrounds the socio-technical and infrastructural dimensions that determine whether such systems can be deployed responsibly. The analysis shows that memory augmentation can improve personalization and intent alignment, but it introduces new failure modes related to stale memories, privacy leakage, feedback loops, and resource consumption. These challenges require coordinated architectural and organizational safeguards, including memory access controls, audit frameworks, lifecycle policies, and evaluation protocols that extend beyond predictive accuracy. The paper concludes by outlining open research directions and policy considerations for sustainable and fair deployment of memory-augmented recommendation systems.

## Keywords

large language models, memory augmentation, e-commerce recommendation, retrieval-augmented generation, user modeling, system governance, fairness, sustainability.

## 1. Introduction

Large language models have begun to reshape the design of recommendation systems by introducing richer semantic understanding of user intent, product catalogs, and interaction contexts than earlier collaborative filtering and content-based approaches typically allowed [1]. Their core representational capability derives from transformer architectures that model long-range dependencies in sequence data [2]. These models also exhibit notable few-shot learning behavior, meaning that they can adapt to new tasks from a small number of examples without extensive task-specific training [3]. In e-commerce settings, however, few-shot behavior alone is insufficient because user preferences evolve continuously and are often expressed through sparse, ambiguous, and context-dependent signals. Static parametric models lack a persistent representation of a user that can be updated through ongoing interaction. Retrieval-augmented generation offers one architectural response by allowing a language model to condition on external knowledge sources at inference time [4]. This paradigm aligns naturally with recommendation scenarios, where product catalogs, purchase histories, and session activities can be treated as external memory. Early neural collaborative filtering models established the value of learning nonlinear user-item interactions [5], while sequential recommenders demonstrated that modeling ordered behavior is essential for capturing preference dynamics [6]. Memory-augmented large language models combine these lineages by integrating external user memory with strong language understanding and sequence modeling.

A central challenge in this integration is domain adaptation. General-purpose language models encode broad linguistic and world knowledge, but they have not necessarily internalized the structured inventory systems, price sensitivities, brand hierarchies, and purchase cycles that characterize electronic commerce. Domain-adaptive pretraining and continued training on in-domain corpora can reduce this mismatch [7]. Nevertheless, adaptation alone does not solve the problem of persistent user state. Even a well-adapted model must receive user-specific information through prompts, retrieval, or fine-tuning. The architectural question is therefore how to arrange long-term user memory relative to the base model, the retrieval system, and the decision layer. Deep learning recommendation systems already maintain multiple representations of users and items, and they rely on careful trade-offs between generalization, memorization, and refresh frequency [8]. Memory-augmented language models inherit these trade-offs while adding the complexity of natural language generation and interpretative reasoning. Consequently, their deployment requires a systems perspective that accounts for data governance, infrastructure cost, robustness, and policy constraints rather than a purely algorithmic perspective.

This paper examines memory-augmented large language models for adaptive e-commerce recommendation from a system-level viewpoint. It considers the architectural foundations of memory integration, the lifecycle of memory entries, structural trade-offs in retrieval and generation, and the implications for deployment at scale. It further addresses fairness, privacy, sustainability, and regulatory concerns that arise when persistent user models are coupled with generative capabilities. The aim is to provide an interdisciplinary synthesis that connects technical architecture with organizational and societal considerations.

## 2. Architectural Foundations of Memory-Augmented Recommendation

Memory-augmented recommendation architectures can be understood as compositions of several interacting subsystems: a base language model, a memory encoder, a retrieval index, a fusion layer, and a policy controller. Deep learning based recommendation systems have

historically separated user representation, item representation, and interaction modeling [8]. Memory augmentation extends this separation by externalizing user state into a store that can be queried and updated independently of the model weights. A memory-based perspective on query matching and intent modeling has been proposed to improve search efficiency through user-specific memory systems [9]. In such a design, memory entries may encode past queries, product views, purchase events, rejection signals, and inferred preferences. The language model then treats retrieved entries as contextual evidence when generating a recommendation, explanation, or follow-up question. This approach is structurally different from fine-tuning, in which user information is compressed into model parameters. External memory permits more interpretable inspection, more granular deletion, and more flexible updates, but it also introduces retrieval latency, storage overhead, and consistency requirements.

The structural trade-offs in memory-augmented recommendation center on what should be stored, when it should be retrieved, and how it should influence generation. Storing every interaction maximizes fidelity but creates a memory system that is difficult to search and may contain outdated or contradictory signals. Storing only summarized preferences reduces storage and latency but risks losing the contextual detail needed for accurate recommendations. The retrieval policy therefore becomes a governance mechanism as much as a technical component. For example, a memory controller may decide that recent session activity has greater relevance than older purchase history, or that high-value product categories deserve distinct memory treatment. These decisions encode assumptions about user behavior and business priorities. When language models are treated as autonomous reasoning engines, there is a risk of treating their outputs as neutral or fully informed, but the underlying retrieval and memory selection processes impose significant editorial control over what the model can know. The broader literature on large language models cautions against overestimating the coherence and neutrality of systems that are ultimately shaped by training data and architectural constraints [10]. In e-commerce, this caution extends to the memory layer itself.

A further architectural consideration is the separation between parametric knowledge and nonparametric memory. Parametric knowledge includes product category knowledge, common shopping language, and general recommendation patterns learned during training. Nonparametric memory includes the user-specific events and preferences held in the external store. The fusion layer must determine how to combine these sources when they conflict. If a user has repeatedly searched for organic products but the parametric model associates a category with conventional brands, the memory system should provide corrective evidence. However, over-weighting memory can reproduce historical biases and limit exposure to new alternatives. Under-weighting memory can make the system feel generic and unresponsive. These trade-offs are not resolved by scaling the language model alone. They require careful design of the controller that balances parametric priors, retrieved memory, and immediate session context. Different e-commerce platforms may adopt different balance points depending on their catalog size, user base, and strategic objectives.

### **3. Memory Lifecycle and Adaptive User Modeling**

The lifecycle of a memory entry is more complex than the lifecycle of a model parameter. Memory entries are created from user interactions, refined through aggregation, retrieved under specific conditions, and eventually deprecated or deleted. This lifecycle must reflect both the temporal dynamics of consumer behavior and the regulatory requirements associated with personal data. In sequential recommendation, the recency and order of interactions often

carry predictive signal [6]. Memory-augmented language models can exploit this signal by assigning temporal weights to retrieved entries or by organizing memory into session-level and long-term layers. Retrieval-augmented generation provides an established mechanism for bringing external evidence into the generation process [4]. However, e-commerce memory systems differ from open-domain retrieval because the evidence is highly personal and commercially consequential. A memory entry that captures a rejected product or a return event can prevent repeated poor recommendations, but an entry that captures a transient interest may incorrectly suppress diversity if retained too long.

Adaptive user modeling requires mechanisms for memory consolidation, conflict resolution, and forgetting. Consolidation transforms raw interaction traces into higher-level preference statements, such as inferred price sensitivity or category affinity. Conflict resolution addresses the fact that users may express contradictory signals across sessions, devices, or household accounts. Forgetting allows the system to remove or down-weight memories that no longer reflect current intent. These operations must be tied to explicit policies rather than implicit model behavior. For example, a platform might retain detailed session memory for a limited period while preserving aggregated preference vectors for longer durations. Such policies create a layered memory architecture that balances responsiveness with privacy and storage constraints. The interpretation of user intent is therefore not a single inference but a continuous negotiation between fresh observations and accumulated history.

Memory lifecycle decisions also influence the stability of recommendations. A system that updates memory too aggressively may overreact to a single click and create volatile recommendations. A system that updates too slowly may miss significant life events or seasonal changes. Recommendation stability is important for user trust, especially when recommendations are accompanied by natural language explanations. If the explanation references memory entries that have been deleted or altered, the system may appear inconsistent. This coupling between memory lifecycle and explanation coherence is an underexamined risk. In adaptive e-commerce systems, explanations often function as commitments about why a product was recommended. When the underlying memory changes, the historical justification may no longer be reconstructable. The governance implication is that memory systems should maintain sufficient provenance metadata to support audit and explanation over time.

#### **4. Robustness, Fairness, and Governance**

Memory-augmented recommendation systems inherit the fairness and bias challenges of both machine learning and information retrieval. Biases may enter through training data, retrieval priorities, memory update policies, or the language model itself. Research on bias and fairness in machine learning has documented the many mechanisms by which automated decisions can produce disparate outcomes [11]. In e-commerce, such disparities may affect which products receive visibility, which users receive premium alternatives, or how promotional language is framed. Memory systems can amplify or mitigate these disparities depending on how they encode historical behavior. If a memory store captures past engagement produced by a biased interface, then recommendations generated from that memory may perpetuate the same bias. The coupling between retrieval and generation makes this feedback loop especially difficult to detect because the final recommendation can appear to be grounded in user-specific evidence even when the evidence itself is the product of structural inequality. Language models also carry broader ethical and social risks, including the potential to produce harmful, manipulative, or exclusionary content [12]. These risks are relevant in e-commerce when

models generate persuasive product narratives or attempt to infer sensitive attributes from browsing patterns.

Fairness in memory-augmented recommendation cannot be reduced to a single metric or a single stage of the pipeline. It must be addressed through a combination of representational safeguards, retrieval audits, and governance mechanisms. Representational safeguards include ensuring that product descriptions and training corpora do not encode biased associations. Retrieval audits examine whether memory selection privileges certain user groups or product categories. Governance mechanisms establish responsibilities for monitoring and correcting unfair outcomes. The memory layer introduces additional governance requirements because it creates a persistent digital representation of a user. This representation can become part of downstream decisions beyond recommendation, such as credit scoring, insurance targeting, or dynamic pricing. Even if the current deployment is limited to product recommendation, the existence of a detailed user memory raises questions about secondary use and data minimization. Organizations must therefore define the purpose of memory augmentation explicitly and restrict its use to that purpose.

Robustness is also a structural concern. Memory-augmented systems must handle noisy retrieval, adversarial product descriptions, and manipulation attempts by sellers. A retrieval index can be polluted by content that is designed to exploit the model, much like search engine optimization affects web search. When such content enters memory and influences generation, the system may produce recommendations that appear trustworthy but are commercially manipulated. Defenses include source verification, retrieval provenance, and periodic memory reconciliation. However, these defenses increase latency and operational complexity. The trade-off between robustness and responsiveness must be evaluated at the system level, taking into account adversarial incentives that are specific to e-commerce marketplaces.

## **5. Sustainability and Energy Trade-Offs**

The computational demands of large language models have drawn attention to their energy consumption and environmental impact. Training and serving such models require significant resources, and the implicit costs are often distributed unevenly across communities and regions [13]. Memory augmentation adds further demands because it requires not only model inference but also retrieval, memory encoding, and periodic consolidation. The energy footprint of a memory-augmented recommendation system is therefore not limited to the language model parameters. It includes the storage infrastructure, the retrieval index, the update pipeline, and the network traffic associated with moving memory entries across services. Sustainability assessments must account for the full serving environment rather than the model alone. Work on carbon emissions in large neural network training demonstrates the importance of measuring hardware, electricity mix, and operational lifespan [14]. For recommendation systems that run continuously at high request volumes, serving energy can exceed training energy over time.

System designers can reduce energy consumption through several structural mechanisms, including smaller specialized models, memory summarization, and selective retrieval. Instead of querying a full memory store for every recommendation, the system may retrieve only when the session context indicates high uncertainty or when the cost of retrieval is justified. Memory consolidation can compress interaction histories into compact representations, reducing both storage and retrieval energy. Caching frequent memory reads can further reduce repeated computation. However, these optimizations interact with fairness and

accuracy. Aggressive summarization may discard signals that matter for minority users or niche products. Selective retrieval may improve average efficiency while increasing variance in recommendation quality. The sustainability objective is therefore not independent of the fairness objective. Both require careful allocation of computational resources across users and use cases.

A broader sustainability perspective includes organizational and economic dimensions. The decision to deploy memory-augmented models should be justified by measurable improvements in user experience, conversion, or satisfaction, not merely by technological novelty. Many e-commerce recommendation tasks may be served adequately by simpler methods with lower resource use. Memory augmentation is most valuable when recommendations require long-term personalization, contextual reasoning, or explanation generation. In such cases, the energy investment may be justified if it reduces failed recommendations, returns, or excessive search costs. Lifecycle assessment should therefore consider not only model training and inference but also the downstream effects on consumer behavior and logistics. These considerations connect sustainability to business strategy and societal impact.

## **6. Infrastructure, Serving, and Deployment Trade-Offs**

Deploying memory-augmented large language models in production requires an infrastructure that can manage high throughput, strict latency budgets, and dynamic memory access patterns. Sparsely-gated mixture-of-experts architectures have demonstrated that large model capacity can be combined with conditional computation to reduce serving cost [15]. Such approaches are relevant when the base language model is large and only a subset of experts must be active for each request. Switch transformers further simplify routing while scaling to very large parameter counts [16]. Transfer learning frameworks such as the text-to-text transformer provide a unified interface that can simplify deployment across multiple recommendation-related tasks [17]. These architectural advances reduce some of the computational burdens associated with large models, but they do not eliminate the memory retrieval bottleneck. In many deployments, the retrieval layer rather than the language model becomes the limiting factor because memory lookup must be performed before generation can begin.

The serving architecture must also support dynamic updates. Unlike static model weights, memory entries change frequently. User interactions arrive continuously, and the memory store must reflect these changes without interrupting service. This requirement introduces challenges related to consistency, versioning, and synchronization. A recommendation generated from stale memory may be suboptimal, but a system that blocks on memory writes may violate latency constraints. Common solutions involve asynchronous updates, eventual consistency, and separation of session state from long-term memory. Session state can be maintained locally with fast access, while long-term memory is updated through a background pipeline. This separation is conceptually attractive but operationally complex because session state and long-term memory must be reconciled at decision time. The policy controller must decide how much weight to assign to each source.

Infrastructure choices also shape the economics of deployment. The cost of serving large language models has decreased through techniques such as distillation, quantization, and specialized hardware, but memory augmentation introduces additional storage and retrieval costs that scale with the number of users and the richness of the memory schema. Organizations must decide whether to centralize memory in a shared service or distribute it across edge nodes. Centralized memory simplifies governance and audit but increases

network latency and data concentration risk. Distributed memory can improve responsiveness and resilience but complicates consistency and deletion. These trade-offs are particularly important in cross-border e-commerce, where data residency requirements may force memory partitioning across jurisdictions. The deployment design must therefore align with regulatory constraints, latency requirements, and organizational capacity for operational management.

The engineering of memory-augmented systems also requires continuous evaluation. Offline benchmarks based on historical click-through data may not capture the adaptive effects of memory updates. Online evaluation must monitor not only accuracy but also latency, staleness, retrieval quality, and failure rates. If recommendations are generated with natural language explanations, evaluation should include the coherence and fidelity of those explanations. Large language models trained on code and instruction-following corpora have shown that task-specific tuning can improve reliability in constrained settings [18]. Instruction tuning with human feedback further increases alignment with user intent [19]. These techniques can be applied to recommendation-specific prompts and memory summaries, but they introduce additional data collection and labeling requirements. The overall system must be managed as a continuously evolving socio-technical infrastructure rather than a static model artifact.

## **7. Policy Implications and Long-Term Governance**

Memory-augmented recommendation systems raise policy questions about user autonomy, transparency, and the boundaries of personalization. The concept of the filter bubble has long warned that personalized systems can narrow the information environment by reinforcing prior preferences [20]. Memory-augmented language models intensify this risk because they can generate persuasive natural language that appears attuned to an individual user. The user may not know which memory entries influenced a recommendation or whether those entries are accurate. Transparency mechanisms such as memory disclosure, explanation generation, and user-facing controls become essential. Sequential recommendation models such as bidirectional encoder representations for sequential data [21], recurrent session models [22], and self-supervised sequential recommenders [23] show that historical interaction modeling is powerful, but their internal representations are often opaque. Memory-augmented language models can be more interpretable if the memory store is designed for inspection, but interpretability is not automatic. Retrieval-augmented generation for large language models has expanded rapidly [24], yet many deployments remain difficult to audit because retrieval provenance is not exposed to users or regulators.

Policy frameworks should require that memory systems be subject to data protection principles including purpose limitation, storage limitation, and data subject access. Users should have the ability to view, correct, and delete memory entries that influence their recommendations. However, deletion is not straightforward when memory entries have been aggregated into preference vectors or embedded representations. The system must document how deletion propagates through derived artifacts. If an aggregated representation cannot be fully decomposed, the organization may need to retrain or recompute the affected model states. Governance mechanisms must therefore include technical deletion procedures, not just legal commitments. The memory layer should be designed with data minimization in mind, so that the minimum necessary information is retained for the shortest appropriate period.

Long-term governance also involves market and regulatory dynamics. Recommendation systems can shape consumer behavior at scale, affecting price discovery, product visibility, and competition among sellers. Memory-augmented systems that prioritize certain products or

sellers through personalized reasoning may create new forms of information asymmetry. Sellers may attempt to manipulate memory systems through sponsored content, fake reviews, or adversarial product descriptions. Regulatory oversight may need to address not only consumer privacy but also fair competition. The use of persuasive language generation in commerce also raises concerns about manipulation, particularly when models exploit inferred emotional states or vulnerabilities [25]. Governance should therefore include ethical review of deployment contexts, monitoring for manipulative patterns, and safeguards for vulnerable populations.

The future evolution of memory-augmented recommendation is likely to involve more autonomous memory update policies, richer multimodal memory stores, and deeper integration with conversational shopping interfaces. These developments will increase the stakes for system-level governance. Organizations that deploy such systems should establish cross-functional oversight involving engineering, legal, product, and policy teams. They should also invest in audit infrastructure that can trace a recommendation back to the memory entries and policy decisions that produced it. The goal is not to eliminate personalization but to ensure that personalization operates within transparent and accountable boundaries.

## 8. Conclusion

Memory-augmented large language models offer a promising direction for adaptive e-commerce recommendation by combining semantic reasoning with persistent user-specific memory. This paper has examined the architectural foundations, memory lifecycle, robustness, fairness, sustainability, deployment, and policy implications of such systems. The analysis indicates that the primary challenges are no longer purely model-centric. They involve the design of memory controllers, retrieval policies, provenance mechanisms, and governance structures that can handle dynamic user data responsibly. Memory augmentation improves the capacity to interpret ambiguous intent, but it introduces new risks related to stale evidence, feedback loops, privacy, manipulation, and resource consumption. These risks require coordinated action across technical and organizational boundaries. Future work should develop standardized evaluation protocols for adaptive memory systems, including measures of stability, fairness, energy efficiency, and explanation fidelity. It should also investigate memory deletion, user control, and audit mechanisms under realistic deployment conditions. By treating memory-augmented recommendation as a socio-technical system rather than a standalone algorithm, researchers and practitioners can better align technical innovation with user welfare and societal values.

## References

1. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, Y., Hu, Y., Wei, Y., Liu, Z., Chen, E. (2024). A survey on large language models for recommendation. *World Wide Web*, 27(5), Article 60.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

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. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
5. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. *Proceedings of the 26th International Conference on World Wide Web*, 173–182.
6. Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. *Proceedings of the 2018 IEEE International Conference on Data Mining*, 197–206.
7. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.
8. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), Article 5.
9. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
10. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115.
12. 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*.
13. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
14. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
15. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Proceedings of the 5th International Conference on Learning Representations*.
16. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1–39.
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.

18. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, 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.
20. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.
21. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1441–1450.
22. Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. *Proceedings of the 4th International Conference on Learning Representations*.
23. Zhou, K., Wang, H., Zhao, W. X., Zhu, Y., Wang, S., Zhang, F., Wang, Z., & Wen, J.-R. (2020). S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1893–1902.
24. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
25. McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:2009.06807.