

Personalized Product Search Through Joint Modeling of User Intent, Behavioral History, and Semantic Memory

Ashwin R. Pathak

Department of Computer Science, University of North Texas, Denton, TX, USA.
arpathak@unt.edu

Jan C. Garrett

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
garrett206@ucf.edu

Abstract

Personalized product search operates at the intersection of information retrieval, user modeling, and large-scale machine learning infrastructure. This paper develops a systems-level analysis of joint modeling approaches that integrate three complementary streams: query-time user intent, longitudinal behavioral history, and persistent semantic memory. Rather than proposing a single ranking algorithm, the paper examines architectural trade-offs, data governance requirements, deployment patterns, and policy implications for production-scale personalized retrieval. The discussion addresses how intent representations capture immediate purchase goals, how behavioral history encodes temporal and contextual preferences, and how semantic memory supplies durable conceptual structures that stabilize personalization across sessions. Drawing on retrieval, ranking, transformer, recommender systems, privacy, and fairness literatures, the paper identifies structural tensions between real-time latency, model capacity, privacy preservation, and auditability. It further analyzes feedback loops, concept drift, cold-start conditions, and regulatory constraints. The central argument is that durable value emerges not from isolated ranking improvements but from a governed system of memory consolidation, intent inference, and policy-aware retrieval. The paper closes with a forward-looking research agenda emphasizing energy efficiency, fairness, auditability, and cross-domain transfer.

Keywords

personalized product search, joint modeling, user intent, behavioral history, semantic memory, information retrieval, system governance, large language models.

1. Introduction

Personalized product search has evolved from lexical matching toward user-aware ranking systems that must reconcile immediate query signals with long-term preference evidence. Traditional information retrieval models assumed that relevance could be judged primarily from query-document similarity, but commercial product search quickly revealed that identical queries from different users can imply divergent purchase intentions [1]. Early learning-to-rank formulations demonstrated that clickthrough data could be treated as implicit relevance feedback and used to optimize ranking functions [2]. Subsequent work showed that behavioral information such as dwell time, query reformulations, and purchase sequences provides rich evidence beyond static product descriptors [3]. These developments motivated a

shift from isolated query relevance toward a broader view of the user as a situated actor whose preferences are dynamic and context dependent.

Industrial product search systems nevertheless continue to face structural tensions. Active exploration of the ranking space introduces risk because poor results in commercial settings directly affect revenue and user trust [4]. Personalization models must distinguish short-term task goals from persistent preferences and must decide how much weight to assign to each signal under uncertainty [5]. Automated analysis of interest profiles has shown that even coarse personalization can improve ranking quality, yet profile construction raises questions about data staleness, coverage, and interpretability [6]. Similar lessons emerge from sponsored search, where personalized click prediction must operate under strict latency budgets and adversarial click patterns [7]. These prior strands suggest that no single relevance model can satisfy the demands of modern e-commerce. Instead, systems must jointly represent user intent, behavioral history, and semantic memory.

2. System-Level Framing and Architectural Challenges

A systems-level perspective treats personalized product search as controlled information flow across three interacting subsystems: an intent interpretation layer, a behavioral evidence layer, and a semantic memory layer. Intent interpretation converts the current query and session context into representations of immediate purchase goals. Behavioral history supplies temporal sequences of clicks, carts, purchases, and returns that reveal preference patterns. Semantic memory maintains stable conceptual and relational knowledge about products, categories, attributes, and user preferences over longer horizons. The central architectural question is not whether personalization matters, but how to couple these layers without sacrificing latency, auditability, or model stability.

Coupling decisions determine failure modes. Tightly coupled systems that feed all signals into a single monolithic model can achieve high accuracy in controlled evaluations but are difficult to govern in production because changes in one signal distribution propagate unpredictably. Loosely coupled systems that treat the three layers as separate services with well-defined interfaces are easier to monitor and update, but they may miss deep interactions among intent, history, and memory. Hybrid designs often use shared embeddings and asynchronous memory stores, allowing the query-time path to remain fast while behavioral and semantic signals are updated offline. These architectural trade-offs shape the remainder of this paper.

3. User Intent Representation and Query Understanding

Modern query understanding has moved beyond bag-of-words retrieval toward dense representation learning. Recent work on memory systems for query matching and intent modeling indicates that user-specific memory can improve search efficiency when integrated into the matching pipeline [8]. Transformer architectures provide self-attention mechanisms that allow query words and product metadata to be contextualized jointly [9]. Pretrained bidirectional encoders such as BERT improved query and document representation by capturing deeper lexical and semantic relationships [10]. Dense passage retrieval further demonstrated that dual-encoder models can retrieve relevant passages efficiently at scale [11]. Contrastive training with approximate nearest neighbor negative sampling improved dense retrieval quality without prohibitive computational cost [12]. These advances make it feasible to encode product queries and catalog items in a shared vector space where user intent can be represented as a learned prior.

Query-time intent inference nevertheless remains brittle when queries are short, ambiguous, or novel. A query such as lightweight jacket may correspond to different temperature ranges, activities, or style preferences depending on the user's recent browsing and purchase history. The system must therefore decide how much weight to place on the lexical query, the session context, and the user's long-term profile. Dense encoders can be fine-tuned with click and purchase supervision, but the resulting models may overfit to head queries and fail to generalize to sparse catalog categories. Regularization through contrastive objectives and data augmentation helps, yet the underlying problem is representational: intent is a latent condition that cannot be reduced to query text alone. A well-designed intent layer therefore functions as a probabilistic filter that conditions retrieval on user state while preserving the ability to fall back to lexical matching when confidence is low.

4. Behavioral History Integration and Temporal Dynamics

Behavioral history provides the strongest empirical signal for personalization because it reflects actual choices rather than stated preferences. However, history is heterogeneous and temporally structured. Short-term sessions reveal immediate task goals, while long-term logs reveal durable preferences such as brand affinities, price sensitivity, and category interests. Learning from these signals requires ranking models that can incorporate both short- and long-term behavior without allowing recency to dominate irrelevant context [5]. Preference optimization methods derived from human feedback can help align ranking behavior with implicit satisfaction signals, though they introduce calibration and reward design challenges [13]. Large language models have also demonstrated emergent capabilities that are useful for interpreting sparse behavioral evidence across many product domains [14]. Recommender systems research provides a parallel body of work on neural collaborative filtering and sequence modeling that can inform product search architectures [15].

The temporal dynamics of behavioral history are especially challenging in e-commerce because purchase cycles vary widely across categories. A user may buy a laptop once every several years but purchase household consumables weekly. A uniform decay function over historical events is therefore inadequate. Systems must instead maintain category-specific memory horizons and event types. Clicks may indicate exploration, while carts and purchases indicate commitment, and returns indicate dissatisfaction. Each event type should enter the model with different weight and decay semantics. Furthermore, behavior can be influenced by promotions, seasonality, and external shocks, making attribution uncertain. Robust systems treat behavioral history not as a fixed profile but as a continuously reweighted evidence stream whose uncertainty is propagated into ranking decisions.

5. Semantic Memory Infrastructure

Semantic memory complements behavioral history by storing durable conceptual relationships that do not rely solely on user actions. In product search, this includes category taxonomies, attribute graphs, compatibility constraints, brand hierarchies, and natural language descriptions of product features. Semantic memory allows the system to generalize across superficially different items and to reason about substitutability and complementarity. Structured memory schemas can benefit from advances in code-oriented language modeling, where syntactic constraints and type systems improve the fidelity of stored representations [16]. Transformers and dense retrieval models can populate semantic memory with embeddings that serve as long-term referents for product entities [11]. The memory layer functions as a knowledge substrate that stabilizes personalization when behavioral evidence is sparse or contradictory.

A critical architectural distinction is between parametric memory embedded in model weights and nonparametric memory stored in external indexes. Parametric memory compresses knowledge into transformer parameters, enabling fast inference but making updates expensive and difficult to audit. Nonparametric memory, such as vector databases and knowledge graphs, can be updated incrementally and inspected directly. Most production systems adopt a hybrid approach: parametric encoders generate representations, while nonparametric stores retain product facts, user summaries, and semantic relationships. This separation improves freshness and reduces the operational cost of catalog updates, and it also provides a clearer boundary for compliance and explainability.

6. Joint Modeling and Inference Trade-offs

Joint modeling of intent, behavior, and memory can be achieved through early fusion, late fusion, or learned gating mechanisms. Early fusion concatenates all available features before scoring, which allows the model to capture complex interactions but increases inference cost and can obscure the contribution of each signal. Late fusion computes separate scores for query relevance, behavioral affinity, and semantic compatibility, then combines them with learned or heuristic weights. Learned gating offers a middle path by allowing the model to modulate the influence of each subsystem based on confidence estimates. The structural trade-off is between expressiveness and interpretability. High-capacity joint models often outperform simpler fusion methods but are harder to debug, monitor, and update safely.

Inference latency and memory access are central constraints. Query-time personalization must be completed within tens or hundreds of milliseconds, while behavioral and semantic stores may contain millions of user and product records. Caching frequently accessed user profiles reduces latency but introduces staleness. Precomputed user embeddings allow fast ranking but cannot capture session-specific intent shifts. The most robust architectures separate candidate retrieval from ranking. A lightweight retrieval stage uses approximate nearest neighbor search over product embeddings conditioned on the query and a coarse user summary, while a heavier ranking stage re-ranks candidates using full behavioral and semantic features. This staged design permits real-time personalization without allowing the entire joint model to operate over the full catalog.

7. Deployment, Scalability, and Infrastructure Governance

Deploying a joint personalized product search system requires coordinated infrastructure for model training, serving, and continuous evaluation. User-specific data is often distributed across devices, regions, and legal jurisdictions, which motivates decentralized training approaches. Federated learning enables model updates without centralizing raw behavioral logs [17]. However, federated training introduces heterogeneity across clients and increases coordination complexity. Differential privacy provides formal guarantees for limiting the leakage of individual user information during model training [18]. These techniques are not merely operational add-ons. They shape what can be stored in semantic memory and how behavioral history may be used for personalization.

Infrastructure governance also includes feature versioning, model rollback, and online experimentation. Because personalized ranking relies on latent representations that shift when models are retrained, version compatibility between encoders and memory stores must be maintained carefully. Embedding spaces that drift across model versions can produce inconsistent rankings and erode user trust. Contractual interfaces among the intent, history, and memory services help isolate change. Monitoring must track not only aggregate purchase

conversion but also degradation in protected user segments and catalog coverage. These governance mechanisms are essential for sustaining personalized search at scale.

8. Fairness, Privacy, and Policy Implications

Personalization systems can amplify existing inequalities if behavioral history encodes historical discrimination or if semantic memory inherits biased product representations. A large body of work has documented how machine learning models can encode and reproduce bias [19]. Fairness in personalized search requires choices about whether to equalize ranking exposure, outcome rates, or user satisfaction across groups [20]. For example, a model that heavily weights past purchases may recommend fewer expensive items to users from lower-income groups, even when those users have equivalent purchase intent. Such behavior can be commercially rational in the short term but harmful in the long term, reinforcing exclusion and limiting market development.

Privacy regulations increasingly constrain how behavioral histories may be collected, retained, and used for ranking. Techniques such as data minimization, purpose limitation, and user deletion rights must be designed into the retrieval architecture rather than retrofitted. Differential privacy, federated learning, and on-device inference reduce the centralization of sensitive data but may also reduce personalization accuracy. There is therefore a policy-relevant trade-off between retrieval quality and privacy. Product search systems should provide users with transparent controls over what history is remembered and what types of inference are permitted. Audit logs and external review mechanisms can improve accountability without requiring the exposure of proprietary model details.

9. Sustainability and Operational Robustness

The computational footprint of personalized search has grown with transformer-based encoders and dense retrieval indexes. Training large language models consumes substantial energy, and serving personalized models at scale multiplies this cost [21]. Sustainability considerations therefore influence architectural choices. Dual-encoder retrieval is more energy-efficient than cross-encoder scoring over all candidates, distillation can reduce model size, and caching can lower repeated inference. However, efficiency gains must be balanced against ranking quality and fairness. Continuous retraining pipelines also consume energy, requiring careful scheduling and data selection to avoid needless computation.

Operational robustness depends on resilience to catalog churn, concept drift, and adversarial behavior. Product catalogs change rapidly with seasonal inventory, price fluctuations, and new launches. Behavioral history may be polluted by clickbait, bot traffic, and promotional distortions. Semantic memory must be refreshed without catastrophic forgetting of stable category knowledge. Retrieval-augmented generation has been proposed as a mechanism for combining parametric knowledge with external retrieval, which can improve freshness and reduce the need for frequent full retraining [22]. Such techniques also enable personalized search systems to handle emerging product types and long-tail queries more gracefully.

10. Future Research Directions

Several directions merit further research. Cross-domain transfer of semantic memory could help platforms personalize in sparse domains by borrowing durable preference representations from related categories, though this raises new privacy and interpretability challenges. Causal inference methods might distinguish preference-driven behavior from exposure-driven behavior, reducing the feedback loop bias that arises when recommendations shape future

user choices. Lifelong personalization models that consolidate behavioral events into semantic memory over time could reduce the need to retain raw interaction logs, improving privacy and storage efficiency.

Another important direction is the development of auditable personalization contracts that specify what data may influence ranking, how uncertainty is propagated, and how users can contest or correct inferred preferences. These contracts would bridge technical systems design and regulatory governance. Evaluation protocols must move beyond aggregate click-through and conversion metrics toward measures of long-term user trust, supplier fairness, catalog discovery, and environmental cost. Evaluation protocols must move beyond aggregate click-through and conversion metrics toward measures of long-term user trust, supplier fairness, catalog discovery, and environmental cost. Only when these criteria are treated as first-class system objectives can joint modeling avoid externalizing social and environmental harms. Such measures require logging not only displayed rankings and subsequent clicks but also counterfactual exposure patterns. Because personalized product search influences which items users see, the data used to train future intent and memory models are shaped by prior ranking decisions. This creates a self-reinforcing loop in which historical behavior may merely reflect past system choices rather than independent user preference. Causal inference offers a way to separate compositional preference from exposure-driven behavior, but its deployment in large-scale search remains nascent and computationally demanding.

A further structural challenge concerns the consolidation of behavioral history into semantic memory. If raw interaction logs are retained indefinitely, privacy risks and storage costs increase. If history is aggressively summarized into memory vectors, individual events lose interpretability and users lose the ability to challenge particular inferences. A promising pathway is the development of lifelong personalization models that consolidate events incrementally into editable semantic representations. Such systems could satisfy retention minimization requirements while preserving personalization quality. However, consolidation introduces the risk of irreversible distortion when new evidence is inconsistent with older preferences. Mechanisms for time-aware conflict resolution, user-visible preference explanations, and selective forgetting therefore represent important research priorities.

Cross-domain and cross-market transfer also require investigation because semantic memory built in one product category may encode assumptions that do not transfer cleanly to another. Durable preferences for compact and lightweight devices may generalize from electronics to travel accessories, but category taxonomies and attribute vocabularies differ. Joint modeling systems that learn shared user representations across domains can improve cold-start performance in sparse categories, yet they must also respect local cultural and regulatory norms. Federated and distributed architectures may allow cross-market learning without transporting raw behavioral data, but heterogeneous label distributions and concept drift across geographies remain open problems.

Finally, auditability should become a core architectural property. Personalized search systems increasingly operate as private policy engines that allocate visibility and economic opportunity. Regulators, merchants, and users need evidence about why certain products were retrieved and how personal history influenced ranking. This does not require exposing proprietary model weights, but it does require versioned feature logs, interpretable fusion layers, and documented consent configurations. Such infrastructure can support the emergence of personalization contracts that specify what data may influence ranking, how uncertainty is propagated, and how users can contest or correct inferred preferences. The

convergence of these technical and governance mechanisms will determine whether joint modeling of intent, history, and memory becomes a sustainable foundation for next-generation product search.

11. Conclusion

Personalized product search is best understood not as a single ranking model but as a governed socio-technical system that coordinates three distinct forms of information. Query-time intent bounds the user's immediate task, behavioral history supplies empirical evidence about preferences, and semantic memory provides durable conceptual structure. The value of joint modeling lies in managing the interactions among these sources without allowing any one to dominate unchecked. Architectural choices such as early fusion, late fusion, and learned gating are not neutral implementation details; they shape inference latency, interpretability, update cost, and the ability to enforce fairness constraints. A staged retrieval and ranking architecture remains the most practical compromise for real-time commerce, but it must be paired with asynchronous memory consolidation and user-level policy enforcement.

The analysis in this paper shows that governance and technical design are inseparable. Privacy mechanisms such as federated learning and differential privacy constrain what can be remembered and how personalization may be learned. Fairness interventions must address feedback loops in which historical behavior is partly an artifact of prior ranking decisions. Sustainability requires balancing transformer-based semantic encoding against the energy cost of continuous retraining and dense retrieval. Operational robustness demands resilience to catalog churn, adversarial traffic, and concept drift. These pressures are structural rather than incidental, and they become more acute as catalog scale and user expectations grow.

Future progress therefore depends on treating personalized product search as an evolving research frontier where retrieval quality, civil rights, environmental impact, and system auditability are jointly optimized. Cross-domain memory transfer, causal preference modeling, lifelong consolidation, and auditable personalization contracts represent promising directions. The required memory systems must be efficient, editable, and respectful of retention limits. The systems that eventually succeed will combine strong relevance with governed memory, transparent inference, and sustainable infrastructure.

References

1. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
2. Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142.
3. Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19–26.
4. Hofmann, K., Whiteson, S., & de Rijke, M. (2011). Balancing exploration and exploitation in learning to rank online. *Proceedings of the 33rd European Conference on Information Retrieval*, 251–263.
5. Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the impact of short- and long-term behavior on search personalization.

Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 185–194.

6. Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. Proceedings of the 15th International Conference on World Wide Web, 727–736.
7. Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. Proceedings of the 16th International Conference on World Wide Web, 521–530.
8. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
9. 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.
10. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.
11. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781.
12. Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., & Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. Proceedings of the International Conference on Learning Representations.
13. 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., 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., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.
15. 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.
16. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de 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.
17. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

18. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.
19. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 4349–4357.
20. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., & Goodrow, C. (2019). Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2212–2220.
21. 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.
22. 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.