

Personalized Conversational Search with Large Language Models for Long-Term User Preference Modeling

Rrun Kapoor

Department of Computer Science, University of North Texas, Denton, TX, USA.

kapoor1976@unt.edu

Abstract

The design of personalized conversational search systems is undergoing a fundamental shift as large language models become central components of query interpretation, response generation, and retrieval planning. Although these models provide strong semantic generalization, they do not automatically maintain durable, user-specific knowledge over long horizons. This paper develops a system-level analysis of large language model based conversational search architectures that incorporate long-term user preference modeling as an explicit infrastructure concern. It examines the decomposition of user knowledge into episodic interaction state and persistent preference memory, and it discusses how retrieval-augmented generation, dense retrieval, and reranking can interact with governed user models. The paper highlights architectural trade-offs in memory placement, representation granularity, update frequency, latency, index freshness, and forgetting. It then addresses evaluation weaknesses in offline metrics for personalization, the delayed nature of preference feedback, and the challenges of aligning user models with privacy, consent, auditability, and fairness requirements. Through cross-domain comparisons with recommendation and enterprise knowledge systems, the paper argues that sustainable personalization depends less on model scale than on the organizational and technical capacity to reconcile conflicting preference signals, support correction, and maintain audit trails. The discussion concludes with policy implications for multi-tenant deployments under emerging artificial intelligence governance frameworks.

Keywords

conversational search, large language models, user modeling, personalization, retrieval-augmented generation, algorithmic governance.

1. Introduction

Search engines have traditionally relied on static lexical indexing and statistical ranking mechanisms to produce results from query terms [1,2]. These foundations enabled large-scale information access, but they treated each query largely as an independent event. The evaluation of such systems has been formalized through graded relevance judgments and cumulative gain measures [3], while implicit behavioral signals from clickthrough data have been used to optimize ranking from observed interaction patterns [4]. Personalization research has long recognized that user interests persist across sessions and that durable preferences must be combined with short-term query context [5,6]. The emergence of transformer architectures [7] and pretrained bidirectional language models such as BERT [8] allowed matching to become more semantic. Large language models [9] extended this trajectory by enabling generative interpretation of complex intents, and instruction alignment techniques

[10] made these models more responsive to user goals. Retrieval-augmented generation [11] connects generative capabilities to external evidence, while conversational AI research [12] addresses multi-turn interaction, dialogue state, and response coherence.

Despite these advances, long-term personalization remains structurally underdeveloped in many conversational search systems. Large language models can infer immediate intent from a conversational turn, but they do not automatically persist the outcomes of repeated interactions, resolve contradictory preferences, or distinguish stable user interests from transient contextual signals. A system-level view is therefore necessary. The central problem is not merely the selection of an appropriate language model or ranking algorithm, but the design of an infrastructure that can acquire, represent, update, and govern user preference knowledge across time. This paper develops such a perspective by analyzing architectural trade-offs, memory lifecycle considerations, evaluation limitations, deployment constraints, fairness risks, and policy requirements.

2. From Query Context to Durable Preference Models

Traditional personalization in search often relied on aggregate profiles, query histories, and manually engineered features that were external to the ranking process. These approaches tended to be brittle because they separated user modeling from semantic matching. Contemporary neural retrieval and generative systems create new opportunities for tighter integration. Dense passage retrieval [13] represents both documents and user memory segments as dense vectors, allowing semantic matching without exact lexical overlap. Zero-shot dense retrieval methods [14] further reduce the dependence on task-specific relevance labels, which is especially important for user preference modeling because explicit relevance judgments about a particular user are scarce. Joint retrieval and reranking methods [15] then improve end-to-end quality by coordinating different stages of the candidate generation and scoring pipeline.

The challenge is that conversational search adds a recursive layer of context. A user may express a request in one turn, refine it in the next, and reference a preference stated several sessions earlier. Episodic dialogue context can be represented as a working memory that supports immediate reasoning, but durable preferences require a separate memory substrate with different update dynamics. If this separation is missing, the system may overfit to the most recent interaction or fail to recall stable long-term interests. Conversely, if the durable memory is too rigid, the system cannot adapt when a preference genuinely changes. This tension between stability and plasticity is one of the central structural issues in long-term personalization.

3. System Architecture for Memory-Centric Conversational Search

A memory-centric architecture for personalized conversational search includes several interacting components: a query understanding module, a dialogue state tracker, a preference memory store, a retrieval module, a generative planner, and a governance layer. Recent proposals for large language model based user memory systems suggest that query matching and intent modeling can be improved by maintaining persistent user-specific memory representations [16]. Earlier personalization research similarly argued that automated analysis of interests and activities can improve search relevance without requiring explicit user effort [17]. These lines of work converge on the idea that user memory should be treated as a first-class system component rather than an afterthought.

The relationship between the language model and the memory store can be designed in several ways. One design treats memory as an external document collection that is dynamically retrieved and prepended to the language model context. This approach preserves model generality and simplifies inspection because the memory segments remain textual or semi-structured. Another design encodes preferences into vector representations that are used as query prefixes or conditioning signals. This can reduce inference cost, but it makes auditability more difficult because the influence of a preference cannot be easily traced. A third design fine-tunes the language model on user-specific interaction traces. Although this can produce fluent personalization, it raises serious concerns about data isolation, model update cost, and the risk of leaking preferences across users.

The external memory design is often preferable in multi-tenant search deployments because it separates user-specific state from shared model parameters. It also enables differential refresh: the retrieval index can be updated without retraining the generative model. However, the architecture must then solve a hard problem of memory selection. Not every previous interaction is relevant to the current query, and retrieving too many memory fragments introduces noise, increases latency, and dilutes the generative response. The system therefore requires a retrieval stage that can rank memory segments according to their likely usefulness for the current conversational state.

4. Lifecycle of Long-Term Preference Signals

Long-term user preference modeling is not a one-time classification task. Preferences emerge gradually from implicit behavior, explicit corrections, and changing life circumstances. The lifecycle includes signal acquisition, representation, consolidation, retrieval, reconciliation, and forgetting. Implicit signals such as clicks, dwell time, query reformulation, and task completion are abundant but noisy [4]. Explicit signals such as saved items, stated preferences, or corrective feedback are sparse but more reliable. A robust system must combine these signals without assuming that their strengths are equal.

One useful distinction is between stable preferences and episodic interests. A user who consistently searches for vegetarian recipes has a stable preference that should persist. The same user may search for a particular restaurant during trip planning, which is an episodic interest that should not dominate future recommendations. Systems that collapse these categories into a single profile often produce awkward personalization failures. Long-term personalization research has emphasized the importance of modeling both short-term and long-term behavior rather than aggregating them indiscriminately [5]. Context-aware query suggestion models have likewise shown that recurrent dialogue context can improve next-turn prediction [6].

The reconciliation problem becomes more difficult when preferences conflict. A user may search for luxury goods in one context and budget alternatives in another. The system should preserve both possibilities and condition on the current context rather than resolving the conflict into a single averaged vector. This requires representing preferences as structured knowledge with provenance, confidence, and decay functions. The representation should also support user correction. When a user says that a preference is outdated or mistaken, the system must propagate that correction to future interactions and potentially to already cached retrieval results.

5. Evaluation and Operational Trade-Offs

Evaluating personalized conversational search is more difficult than evaluating non-personalized ranking. Traditional graded relevance judgments [3] assume a relatively stable notion of relevance for a query, but personalization makes relevance dependent on the user. Offline datasets rarely contain sufficient per-user interaction depth to measure long-term effects. Clickthrough data provide implicit feedback [4], but clicks are influenced by presentation order, result quality, and interface effects. Moreover, the value of personalization may appear only after repeated sessions, which makes standard single-session evaluation inadequate.

Operational trade-offs further complicate deployment. Dense retrieval [13] and zero-shot methods [14] require managing vector indexes that must be updated as user memory changes. Joint reranking [15] can improve accuracy but adds computational latency. The system must decide whether memory retrieval happens as a first-pass candidate generation step or as a post-hoc reranking stage. Running memory retrieval as part of the initial query expansion increases candidate diversity but can introduce irrelevant context. Running it later preserves precision but may miss serendipitous connections. These choices interact with index architecture, sharding, and refresh frequency.

Latency constraints are particularly acute in conversational interfaces because users expect low response times and incremental updates. A memory-heavy architecture may increase model input length, which can raise inference cost and degrade throughput. Techniques such as caching, summarization, and hierarchical memory compression can mitigate these costs, but they introduce additional trade-offs between fidelity and efficiency. Index freshness is another concern. If user memory is updated continuously, the vector index must support incremental writes without blocking retrieval. Batch updates simplify consistency but delay the incorporation of new preferences. The result is an operational landscape in which personalization quality cannot be optimized independently of infrastructure sustainability and cost.

6. Robustness, Fairness, Privacy, and Governance

Large language models carry known risks of reproducing social stereotypes and overgeneralizing from training data [18]. When these models are connected to persistent user memory, new risks emerge. The system may infer sensitive attributes from behavioral traces, encode them into preference representations, and use them in ways that are difficult to audit. Sociotechnical impact assessments of generative AI emphasize that harms often arise from context-specific interactions rather than from model outputs alone [19]. Ethical analyses of language models similarly identify risks related to privacy, discrimination, and undue influence [20]. These concerns are not hypothetical in personalized search because user models can shape information exposure, reinforce existing preferences, and narrow the diversity of results over time.

Privacy is a first-order architectural constraint. Long-term preference memory requires accumulating detailed information about a person, which creates a valuable data asset that may be vulnerable to unauthorized access, profiling, or commercial exploitation. The political economy of such data accumulation has been analyzed through the lens of surveillance capitalism [21]. In that view, personalized systems can become mechanisms for continuous behavioral prediction rather than neutral tools for information access. A defensible design must therefore include data minimization, purpose limitation, and explicit user control. These principles are increasingly reflected in regulatory instruments such as the European Union

Artificial Intelligence Act [22] and risk management frameworks developed by standards bodies [23].

Fairness in personalized conversational search is complicated by the fact that personalization itself depends on differential treatment. The system must distinguish legitimate individual adaptation from harmful discrimination. This distinction cannot be resolved solely by technical metrics. It requires procedural safeguards such as audit logging, model cards, user-facing explanations, and independent review. Governance mechanisms should document how preference signals are acquired, how they are weighted, how long they are retained, and how users can inspect or delete them. Without such mechanisms, long-term user preference modeling may produce highly responsive systems that nevertheless erode user trust and create systemic accountability gaps.

7. Deployment Sustainability and Policy Implications

Deploying memory-centric conversational search at scale requires attention to organizational and environmental sustainability. Fine-tuning separate models for each user is generally impractical and risky. Shared model parameters with isolated memory stores are more sustainable because they limit the number of model variants and reduce duplicated compute. However, the retrieval infrastructure itself consumes energy and requires continuous vector index maintenance. The trade-off between memory richness and inference cost should therefore be evaluated as an infrastructure governance question, not only as a model quality question.

Policy implications extend beyond privacy. User memory systems must be designed for contestability, meaning that users can challenge incorrect inferences and request correction. They must also be designed for portability, so that preferences can be exported or transferred between platforms under suitable safeguards. In enterprise settings, personalization may be governed by organizational policies, collective bargaining agreements, or sectoral regulations. In public-sector search services, there may be strong transparency requirements that limit the use of opaque memory representations. These considerations suggest that the long-term viability of personalized conversational search depends on institutional design as much as on algorithmic innovation.

A forward-looking system architecture should treat user preference memory as a governed knowledge asset. It should maintain provenance for each preference signal, decay functions for obsolescence, confidence estimates for uncertain inferences, and accessible correction mechanisms. It should isolate user-specific state from model parameters and ensure that user data is not inadvertently shared across tenants. It should also support ongoing monitoring for distribution shifts, stereotyping, and feedback loops. By framing personalization as a socio-technical infrastructure, researchers and practitioners can avoid the tempting but fragile assumption that larger models alone will solve the problem of long-term user modeling.

8. Conclusion

Personalized conversational search with large language models presents a distinctive system-level challenge. The generative capabilities of these models improve immediate query understanding, but durable user preference modeling requires a separate memory architecture with careful lifecycle management, evaluation, and governance. This paper has argued that episodic dialogue context and long-term preferences should be treated as distinct but interacting memory layers. It has examined how dense retrieval, zero-shot representations, reranking, and retrieval-augmented generation can be integrated into a memory-centric design

without conflating user-specific state with shared model weights. The analysis has shown that evaluation cannot rely on static relevance metrics alone because personalization unfolds over repeated sessions and depends on delayed feedback.

The structural trade-offs identified here include stability versus plasticity, precision versus recall of memory retrieval, fidelity versus inference efficiency, and personalization versus fairness. These trade-offs cannot be eliminated by a single technical choice. They must be managed through system design, operational planning, and institutional governance. The deployment of such systems should incorporate explicit privacy controls, auditability, consent, correction, and alignment with emerging regulatory frameworks. Future work should develop longitudinal evaluation methods, standardized governance benchmarks, and infrastructure designs that make preference memory inspectable and sustainable. The long-term goal is not merely to make search engines more responsive, but to create conversational search systems that are accountable, robust, and respectful of the people whose preferences they seek to model.

References

1. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
2. Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Pearson Education.
3. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>
4. Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133–142). ACM. <https://doi.org/10.1145/775047.775067>
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. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 185–194). ACM. <https://doi.org/10.1145/2348283.2348312>
6. Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., & Nie, J.-Y. (2015). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management* (pp. 553–562). ACM. <https://doi.org/10.1145/2806416.2806493>
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*. Neural Information Processing Systems Foundation.
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>

9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*. Neural Information Processing Systems Foundation.
10. 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. In *Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation.
11. 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. In *Advances in Neural Information Processing Systems 33*. Neural Information Processing Systems Foundation.
12. Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3), 127–298.
<https://doi.org/10.1561/15000000074>
13. 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.emnlp-main.550>
14. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Neural Information Processing Systems Foundation.
15. Ren, R., Qu, Y., Liu, J., Zhao, W. X., Sheng, Q., Wu, H., & Wang, H. (2021). RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2825–2835). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2021.findings-emnlp.239>
16. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
17. Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (pp. 54–89). Springer. https://doi.org/10.1007/978-3-540-72079-9_2
18. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv*.
<https://arxiv.org/abs/2112.04359>

19. Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé, H., Dodge, J., Evans, R., Hooker, S., Jernite, Y., Luccioni, A. S., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., Vassilev, A., & Zhuo, T. Y. (2023). Evaluating the social impact of generative AI systems in systems and society. arXiv. <https://arxiv.org/abs/2306.05949>
20. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
21. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.
22. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
23. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>