

Conversational Recommendation with Long-Term User Memory for Personalized Travel Planning

Anders Chambers

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
contactanders@unh.edu

Wesley A. Lane

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
lane1979@ku.edu

Seraj Raxena

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
surajsaxena80@oregonstate.edu

Ravide Beck

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
dbeck@uc.edu

Abstract

Travel planning is a complex decision process characterized by interdependent choices, heterogeneous constraints, and deeply personal preferences that evolve over time. Conversational recommender systems have become a promising paradigm for interactive travel assistance because they support iterative preference elicitation, clarification, and explanation in natural language. However, many existing conversational recommenders are designed as session-bound systems that treat each dialogue as an isolated episode. This paper argues that effective personalized travel planning requires the integration of long-term user memory across sessions, enabling the system to maintain stable preferences, contextual constraints, and interaction histories while adapting to changing user circumstances. The paper examines the architectural foundations of memory-augmented conversational recommendation, including modular dialogue management, memory encoding and retrieval, and recommendation policy design. It presents a system-level analysis of structural trade-offs related to scalability, latency, robustness, privacy, fairness, and sustainability. The discussion emphasizes the need to balance recommendation quality with user autonomy, data governance, and computational cost. Travel planning is used as a representative domain because it exposes the tensions between personalization and persistent user modeling, between rich contextual inference and data minimization, and between model capability and operational efficiency. The paper concludes by outlining future research directions for longitudinal evaluation, adaptive forgetting, federated memory architectures, and governance mechanisms that can support trustworthy long-term user memory in conversational recommender systems.

Keywords

conversational recommendation; long-term user memory; travel planning; personalization; dialogue systems; system architecture; governance; fairness; sustainability.

1. Introduction

Travel planning is a cognitively demanding activity in which users must coordinate flights, lodging, local transportation, dining, and activities under constraints such as budget, time, mobility, dietary needs, and subjective aesthetic preferences. Unlike single-item recommendation, travel planning requires reasoning about bundles of choices that interact with one another and are influenced by situational factors such as weather, health, travel companions, and evolving preferences. Conversational recommender systems have emerged as a natural fit for this setting because they allow users to refine their preferences through multi-turn interaction and receive explanations for suggested options [1]. The conversational paradigm has been studied extensively as a way to reduce the uncertainty that arises in static recommender interfaces by eliciting feedback, negotiating constraints, and explaining trade-offs [2, 3]. However, the majority of conversational recommender research has focused on single-session effectiveness, where the system starts with no persistent representation of the user beyond the current dialogue [4]. This limitation is particularly severe in travel planning because user preferences are often stable across trips, while contextual constraints such as destination, season, and travel purpose change frequently.

Long-term user memory offers a way to extend conversational recommendation from isolated interactions to durable personalized assistance. A memory-augmented system can recall that a traveler prefers early morning flights, avoids certain hotel chains, requires vegetarian meal options, or tends to select cultural activities over nightlife. Such memory can reduce the need for repetitive elicitation, improve the coherence of recommendations, and support proactive suggestions. At the same time, persistent user memory introduces significant system-level challenges. Memory can become stale, inconsistent, or overly intrusive. It can encode sensitive attributes that create privacy risks and discriminatory outcomes. It also raises governance questions about user control, transparency, and the right to be forgotten. Deep learning has expanded the capacity of recommender systems to model rich user representations [5], but the deployment of persistent memory in production travel services requires careful architectural and policy choices.

This paper provides a system-level examination of conversational recommendation with long-term user memory for personalized travel planning. The perspective is deliberately interdisciplinary, connecting recommender systems research with dialogue management, infrastructure design, data governance, fairness, and sustainability. The objective is not to propose a single algorithm or model architecture but to analyze the structural trade-offs that arise when long-term memory becomes a first-class component of conversational recommendation. The paper discusses architectural foundations, memory modeling, dialogue policy, deployment constraints, governance mechanisms, evaluation practices, and future research directions. Throughout, travel planning is used as a case domain that reveals the practical and ethical dimensions of persistent personalization.

2. Background and Related Work

Recommender systems have evolved from collaborative filtering and matrix factorization approaches [6, 7] toward more expressive sequence and session models that capture short-term behavioral dynamics [8]. Classical recommendation methods represent users and items through latent factors or similarity structures, but they do not naturally support multi-turn dialogue or explicit preference revision. Session-based recommenders improved the handling of interaction sequences by learning from click and query traces, yet they remain bounded by the current session and do not provide durable user representations. In travel planning, this

session-bound behavior is especially limiting because user preferences often persist across weeks, months, or years, while individual planning sessions may be separated by long periods of inactivity.

Transformer architectures [9] and pretrained language models [10, 11] have transformed the technical landscape for conversational recommender systems by enabling more fluent natural language generation, richer intent understanding, and flexible representation learning. These models can encode dialogue context and generate recommendations in a unified framework, but they do not by themselves solve the problem of long-term memory. Their context windows are finite, and without external memory they cannot reliably retain user-specific knowledge beyond the current dialogue. Conversational recommenders have therefore begun to incorporate knowledge graphs, external databases, and structured preference representations to improve recommendation quality and dialogue coherence [12, 13]. Knowledge-based approaches can model domain constraints such as hotel locations, restaurant cuisine types, and transportation options, which are essential for travel planning. Yet these systems still tend to treat user knowledge as either static or session-local, rather than as a continuously updated memory that supports identity persistence and contextual adaptation.

The integration of long-term memory into conversational recommendation connects to broader research on personal assistants, user modeling, and retrieval-augmented generation. A memory-augmented conversational system must decide what to store, how to represent it, when to retrieve it, and how to use it in dialogue policy. These decisions are not merely technical; they shape the user experience and the institutional responsibilities of the platform. The travel domain adds further complexity because travel preferences are heterogeneous, temporally situated, and often inferred from sparse signals such as past bookings, search patterns, and explicit statements. A robust memory architecture must therefore support uncertainty, conflict resolution, and selective forgetting.

3. Architectural Foundations for Conversational Travel Recommendation

A conversational travel recommender can be organized as a modular system comprising several interacting components: an intent understanding module, a dialogue state tracker, a user memory manager, a candidate recommendation engine, a policy module, and a response generator. This modular decomposition has been discussed in surveys of conversational search and recommendation [14]. The intent understanding module interprets user utterances and maps them to goals such as booking a hotel, changing a constraint, requesting an explanation, or asking for alternatives. The dialogue state tracker maintains the current conversational context, including the user’s expressed constraints and the set of options under consideration. Separating these modules allows system developers to update individual components without retraining the entire pipeline, which is important for production travel services that must respond quickly to new destinations, policies, or inventory changes.

The user memory manager is the component that distinguishes long-term recommendation from session-bound recommendation. It provides an interface between the dialogue state and persistent user representations. In practice, this manager may combine structured profiles, unstructured interaction logs, preference embeddings, and external knowledge about travel contexts. A modular architecture permits the memory manager to enforce access controls and retention policies independently of the recommendation policy, which is important for privacy and governance. It also allows different memory representations to be used for different purposes. For example, a structured profile might store verified dietary restrictions and

passport-related constraints, while an embedding-based memory might capture softer preferences such as aesthetic style or openness to novel experiences.

One important architectural trade-off is the choice between modular control and end-to-end training. End-to-end neural models can learn rich associations between dialogue context, user history, and recommendation outputs, but they are often opaque and difficult to govern. Modular systems provide clearer accountability because memory access, recommendation decision, and language generation are separated. Variational methods have been proposed to reason about user preferences in conversational recommendation by treating preference as a latent variable that evolves during interaction [15]. Such methods can be embedded in a modular architecture to support uncertainty-aware memory updates. The policy module decides when to ask a question, when to recommend, and when to explain. Conversational recommendation research has examined the interplay between system questions and user responses to improve preference elicitation [16]. Deep conversational recommendation architectures have also been developed to combine dialogue generation with item ranking [17]. These architectures can be extended with an explicit memory layer, but the extension must preserve the ability to audit memory usage and to enforce retention constraints.

Memory-augmented architectures for travel planning must also support asynchronous interactions. Users may begin planning on one device, continue on another, and later return after several weeks. This requires the memory manager to handle session resumption, contextual decay, and cross-session continuity. A service-oriented architecture with a dedicated memory service can support these requirements by decoupling memory persistence from the dialogue runtime. The memory service can expose interfaces for reading, writing, expiring, and deleting user-specific information, while the dialogue policy remains responsible for interpreting the current task. This separation improves modularity but introduces network latency and consistency challenges that must be managed at the system level.

4. Long-Term User Memory Modeling

Long-term user memory in travel planning can be conceptualized as a combination of episodic, semantic, and procedural information. Episodic memory captures past travel experiences such as previous visits to a city, hotel stays, and interactions with prior recommendations. Semantic memory captures generalized preferences and constraints such as preferred cabin class, mobility limitations, or interest in historical sites. Procedural memory captures how the user likes to interact, for example whether they prefer concise confirmations or detailed rationale. A well-designed memory model treats these categories as complementary rather than competing. LLM-based user memory systems have been proposed to enhance search efficiency by improving query matching and intent modeling [18]. In the travel context, such memory can help connect a current query such as a desire for a quiet coastal destination with earlier preferences for boutique hotels and low-crowd environments.

The lifecycle of long-term memory involves acquisition, consolidation, retrieval, decay, and contradiction resolution. Acquisition occurs through explicit user statements, implicit behavioral signals, and inferred preferences from prior recommendations. Consolidation transforms raw interaction records into stable user representations that can be retrieved efficiently. Retrieval is context-sensitive, because only a subset of memory is relevant to a given travel task. Decay is necessary to prevent outdated preferences from dominating future recommendations. Contradiction resolution is especially important in travel planning because users may express preferences that vary by travel mode, companion, or purpose. A business

trip may prioritize proximity to a conference venue, while a leisure trip may prioritize neighborhood character. Memory models must therefore preserve contextual conditions rather than collapsing all signals into a single profile.

Representation choices strongly influence system behavior. Structured representations are transparent and auditable, but they may not capture nuanced preferences. Embedding-based representations support similarity retrieval and natural language integration, but they can be difficult to interpret. Hybrid models can store structured constraints alongside learned embeddings and use the structured layer for policy enforcement. For example, a dietary restriction should be treated as a hard constraint, while a preference for modern architecture may be treated as a soft signal. The memory manager should also maintain confidence estimates and provenance for each piece of information. This allows the conversational policy to ask clarifying questions when memory is uncertain or conflicting.

User control is an essential component of long-term memory modeling. Travelers should be able to inspect, correct, and delete memory entries. Memory should also support selective forgetting at different granularities, such as removing a specific trip, a class of preferences, or all historical data. The system should distinguish between memory that is necessary for a current transaction and memory that is used only for personalization. This distinction is not only a privacy requirement but also a design principle that preserves user trust and reduces the risk of overpersonalization.

5. Dialogue Management and Recommendation Policy

Dialogue management in a memory-augmented travel recommender must balance the use of long-term memory with the need for current situational information. The system should use memory to avoid repetitive questions, but it should not assume that stored preferences remain valid in every context. A traveler who typically selects economy flights may choose business class for an anniversary trip. The dialogue policy should therefore treat memory as a prior rather than a deterministic constraint. This perspective aligns with latent preference reasoning in conversational recommendation [15], where the system maintains a distribution over user preferences and updates it as new evidence arrives.

The recommendation policy faces a structural trade-off between exploration and exploitation. Exploitation uses stored preferences to recommend options that are likely to satisfy the user, while exploration asks questions or suggests novel alternatives to discover new preferences. In travel planning, exploration is important because users often seek novelty when visiting unfamiliar destinations, but excessive exploration increases cognitive load and dialogue length. The policy can use memory confidence to decide between these modes. When memory is highly confident, the system can propose a small set of well-matched options. When memory is uncertain or context has changed, the system can ask targeted clarification questions.

Explanation is another critical function of the dialogue policy. Travel recommendations often involve trade-offs among price, location, convenience, and quality. A memory-augmented system can explain a recommendation by referencing relevant historical preferences, such as previous hotel ratings or expressed interest in walkable neighborhoods. This kind of personalized explanation can improve user trust and acceptance, but it also risks revealing too much about what the system remembers. The dialogue manager should therefore support explanation styles that are transparent but not intrusive. For example, the system might state

that an option was selected because it matches a prior preference for quiet areas without exposing the full memory record.

The dialogue policy must also handle multi-constraint negotiation. In travel planning, constraints often conflict. A user may want a central hotel at a low price and a quiet room. The system must help the user understand which constraints can be relaxed and what the consequences of relaxation are. Long-term memory can inform this process by identifying which constraints have been stable over time and which are likely to be flexible. This allows the system to propose compromises that respect deeply held constraints while relaxing transient ones.

6. System-Level Trade-offs in Deployment

Deploying a memory-augmented conversational travel recommender in production requires attention to latency, scalability, consistency, and cost. The memory subsystem must serve retrievals within the time constraints of natural conversation, which typically requires subsecond latency. This is difficult when memory is large and retrieval involves complex similarity computations. Caching frequently accessed memory entries can reduce latency, but it introduces consistency challenges when the same user interacts across devices or when background processes update memory. A dedicated memory service with asynchronous update queues can balance freshness and responsiveness, but it adds operational complexity. System designers must also decide where memory is stored. Centralized storage simplifies retrieval and governance, while edge or client-side storage can reduce network latency and improve privacy, though it limits the richness of server-side recommendation.

The cost of maintaining long-term memory is not negligible. Memory storage, indexing, and retrieval consume computational resources, and the use of large language models for memory interpretation and dialogue generation adds further cost. These costs must be weighed against the value of improved recommendation quality and reduced user effort. In travel planning, where transactions are often high-value and low-frequency, the cost of memory-augmented inference may be acceptable. However, the system should avoid embedding large language models in every memory operation. Selective retrieval and lightweight candidate filtering can reduce cost without sacrificing conversational quality.

Fairness is also a system-level trade-off. Long-term memory can improve personalization, but it can also reinforce historical biases. If a recommender consistently learns that certain users prefer luxury hotels, it may systematically exclude mid-range options or produce unequal treatment across demographic groups. Fairness in recommender systems has been studied in relation to exposure, opportunity, and outcome [19]. In travel recommendation, fairness concerns may arise in the allocation of visibility to local businesses, the pricing of options shown to different users, and the assumptions made about user preferences based on sensitive attributes. The memory architecture must therefore include mechanisms for auditing stored preferences and detecting patterns that may lead to discriminatory outcomes.

Robustness is another concern. Long-term memory can be corrupted by noisy user statements, outdated preferences, or adversarial inputs. The system should validate high-impact memory entries and allow human override. It should also degrade gracefully when memory is unavailable. In a travel emergency, the system must not refuse to recommend a hotel because the memory service is down. System designers should therefore separate mandatory transaction functions from optional personalization. This separation ensures that memory failures do not block core service delivery.

7. Fairness, Privacy, and Governance

The deployment of long-term user memory in travel recommendation is governed by data protection frameworks such as the European Union General Data Protection Regulation [20]. These frameworks impose obligations related to data minimization, purpose limitation, consent, access, rectification, and erasure. A memory-augmented conversational system must be designed to support these obligations at the architectural level. For example, the memory manager should tag each memory entry with its source and purpose so that deletion requests can be executed precisely. It should also support data portability by allowing users to export their stored preferences. In practice, this requires careful schema design and audit logging.

Fairness cannot be treated as a purely technical property of the recommendation model. Fairness and abstraction in sociotechnical systems require attention to how categories are defined, how outcomes are measured, and how interventions are embedded in organizational practice [21]. In travel recommendation, the use of long-term memory may create abstractions that flatten user identity into a set of inferred preferences. These abstractions can be useful for recommendation but may also reproduce stereotypes about travelers based on location, language, or inferred income. Governance mechanisms should therefore include regular audits of memory representations and recommendation outcomes across different user groups. These audits should examine not only aggregate accuracy but also distributional effects such as the visibility of minority-owned businesses or accessibility of recommendations for travelers with disabilities.

Language models used in conversational memory systems can also encode problematic associations from pretraining data [22]. When memory summarization or preference inference is delegated to a large language model, the system may generate characterizations that reflect societal biases. For example, a model might infer that a traveler from a particular region prefers certain cuisines or that a user with a disability is less interested in outdoor activities. These inferences may be factually wrong and ethically harmful. Governance frameworks should require human review for sensitive inferences, especially when memory is used to make decisions that affect access to services. Users should be able to challenge inferred preferences and receive clear explanations of how memory influences recommendations.

Organizational governance is essential because long-term memory systems span multiple teams and functions. Data scientists, product managers, legal staff, and infrastructure engineers must coordinate on retention policies, access controls, and incident response. A memory governance board or similar structure can establish rules for memory decay, consent, and audit. This board should also define escalation procedures for cases in which memory errors cause harm, such as recommending an inaccessible hotel to a traveler with mobility constraints. The technical architecture alone cannot ensure responsible use; it must be embedded in an institutional framework that assigns accountability.

8. Evaluation and Sustainability

Evaluating a memory-augmented conversational recommender requires more than offline accuracy metrics. Standard recommendation metrics such as precision, recall, and ranking quality are informative but insufficient because they do not capture the longitudinal value of memory or the quality of dialogue. Evaluation should include measures of preference elicitation efficiency, conversation length, user satisfaction, memory accuracy, and user control. Travel planning evaluation should also consider task completion under realistic constraints, such as incomplete information, shifting requirements, and multi-device

continuation. Longitudinal user studies are particularly important because the benefits of long-term memory emerge over time. Short-term experiments may show that memory adds latency or complexity without demonstrating its value in reducing future elicitation.

Memory accuracy can be evaluated through simulated user histories and through human assessments of whether stored preferences reflect current user intent. However, accuracy must be balanced against staleness. A system that remembers too much may appear invasive or outdated. Evaluation should therefore include metrics for memory decay and user correction rates. The frequency with which users edit or delete memory entries is a useful operational signal. A high rate of correction may indicate that memory inferences are too aggressive or that the system is overconfident in its stored representations.

Sustainability has become an important evaluation dimension for systems that rely on large language models and dense retrieval infrastructure. Deep learning models in natural language processing can impose substantial energy and carbon costs [23]. A memory-augmented travel recommender should therefore be evaluated not only on recommendation quality but also on computational efficiency. Techniques such as model distillation, selective retrieval, and parameter-efficient fine-tuning can reduce energy use. The memory subsystem should avoid unnecessary embedding computations by using exact matching for structured constraints and by indexing only the most relevant facets of user history. A sustainability-oriented design may favor modular retrieval over large end-to-end generative models for routine memory operations, reserving generative inference for cases where natural language synthesis is essential.

The operational carbon profile of a memory-augmented system also depends on data center location, energy mix, and load patterns. Travel services often experience seasonal peaks that coincide with holiday periods. System architects can reduce environmental impact by scheduling memory consolidation and index maintenance during periods of lower request volume. These choices are not neutral; they affect the cost structure and responsiveness of the service. A sustainable system is one that treats computational resources as a first-class constraint rather than an afterthought.

9. Future Research Directions

Future research should develop memory architectures that support adaptive forgetting and contextual memory retrieval at scale. Existing systems often rely on fixed retention windows or simple decay heuristics, but travel preferences change through life events such as parenthood, retirement, or health changes. A more sophisticated memory model could detect structural shifts in preference and trigger re-elicitation. Research on programmatic memory policies could allow users to specify how their data should be used across different trip types. This would move beyond a binary privacy setting toward a more nuanced model of user consent.

Federated memory is another promising direction. Rather than centralizing all user memory in a single cloud service, federated architectures could store sensitive preferences on user devices while sharing only aggregate or encrypted signals with the central recommender. This approach raises difficult technical questions about retrieval latency, model quality, and cross-device synchronization, but it could improve privacy and user control. Travel planning is a natural testbed because users often manage itineraries across personal devices and third-party services. Interoperable memory standards would allow users to port or revoke memory across platforms, reducing lock-in and increasing trust.

Finally, evaluation methodologies must mature to capture the long-term sociotechnical effects of memory-augmented recommendation. Researchers should develop open benchmarks for longitudinal conversational recommendation that include fairness, privacy, and sustainability metrics. These benchmarks should not reduce evaluation to a single score but should provide multidimensional profiles that reflect the trade-offs faced by system designers. Case studies of deployed travel assistants would also be valuable to understand how memory systems interact with organizational practices, legal constraints, and user expectations.

10. Conclusion

Conversational recommendation with long-term user memory offers significant potential for personalized travel planning by reducing repetitive elicitation, improving coherence, and enabling proactive assistance. However, the integration of persistent memory into conversational systems is not a simple extension of existing session-based recommenders. It requires a fundamental architectural commitment to memory management, uncertainty-aware dialogue policy, and user control. The travel domain illustrates the structural tensions between personalization and privacy, between rich contextual inference and operational efficiency, and between model capability and governance responsibility. A well-designed memory-augmented system must treat memory as an evolving sociotechnical resource rather than a fixed database of preferences. It must support selective forgetting, transparent explanation, and auditability. It must also be evaluated against a broad set of criteria that include accuracy, dialogue quality, fairness, user satisfaction, and environmental impact. As conversational recommender systems become more embedded in everyday travel planning, the governance of long-term user memory will become as important as the algorithms that use it. The central challenge is to design systems that remember enough to be useful without remembering so much that they undermine autonomy, equity, or trust.

References

1. Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5), Article 105. <https://doi.org/10.1145/3453154>
2. Gao, C., Lei, W., He, X., de Rijke, M., & Chua, T.-S. (2021). Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2, 100–126. <https://doi.org/10.1016/j.aiopen.2021.06.002>
3. Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 815–824). ACM. <https://doi.org/10.1145/2939672.2939746>
4. Sun, Y., & Zhang, Y. (2018). Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 235–244). ACM. <https://doi.org/10.1145/3209978.3210002>
5. 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. <https://doi.org/10.1145/3285029>
6. Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (2nd ed., pp. 1–34). Springer. https://doi.org/10.1007/978-1-4899-7637-6_1

7. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37. <https://doi.org/10.1109/MC.2009.263>
8. Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. In 4th International Conference on Learning Representations (ICLR). arXiv. <https://arxiv.org/abs/1511.06939>
9. 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* (pp. 5998–6008). Curran Associates. <https://arxiv.org/abs/1706.03762>
10. 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, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report. <https://openai.com/research/better-language-models>
12. Zhou, K., Zhao, W. X., Bian, S., Zhou, Y., Wen, J.-R., & Yu, J. (2020). Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1006–1014). ACM. <https://doi.org/10.1145/3394486.3403143>
13. Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., & Tang, J. (2019). Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 1803–1813). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1189>
14. Wu, C., Wang, Y., Ma, W., Zhang, W., Chen, Y., & Wang, X. (2022). A survey on conversational search and recommendation: Datasets, methods, and opportunities. *ACM Transactions on Information Systems*, 40(4), Article 77. <https://doi.org/10.1145/3502870>
15. Ren, Z., Tian, Z., Li, D., Ren, P., Yang, L., Xin, X., Liang, H., de Rijke, M., & Chen, Z. (2021). Variational reasoning about user preferences for conversational recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 165–175). ACM. <https://doi.org/10.1145/3404835.3462853>
16. Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)* (pp. 177–186). ACM. <https://doi.org/10.1145/3269206.3271776>
17. Li, R., Ebrahimi Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards deep conversational recommendations. In *Advances in Neural Information*

Processing Systems 31 (NeurIPS 2018) (pp. 9725–9735). Curran Associates.
<https://proceedings.neurips.cc/paper/2018/hash/1f2b325dcda12a68ed498273206d263-Abstract.html>

18. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
19. Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18) (pp. 2243–2251). ACM. <https://doi.org/10.1145/3269206.3272027>
20. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.
<http://data.europa.eu/eli/reg/2016/679/oj>
21. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19) (pp. 59–68). ACM.
<https://doi.org/10.1145/3287560.3287598>
22. 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 (FAccT '21) (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
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.
<https://doi.org/10.18653/v1/P19-1355>