

Large Language Model-Based User Modeling for Intelligent Customer Service and Query Routing

Jinwan Tian

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

jingwen342@unr.edu

Yaoyang Gu

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

yao791@buffalo.edu

Shenghan Liang

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

shenghanl@ku.edu

Abstract

Intelligent customer service systems increasingly depend on accurate and adaptive representations of users to interpret requests, infer intent, and route queries to appropriate resolution pathways. The emergence of large language models has created new opportunities for building richer user models that operate over unstructured conversational evidence, historical interaction traces, and contextual signals. This paper presents a system-level examination of large language model-based user modeling for intelligent customer service and query routing. It analyzes how language model capabilities can be integrated into user profile construction, dynamic intent inference, and routing decisions across heterogeneous service ecosystems. The discussion emphasizes structural trade-offs among latency, accuracy, explainability, privacy, and operational cost. It further addresses robustness under distributional shift, fairness across demographic and linguistic groups, governance constraints, and deployment sustainability. Rather than prescribing a single architecture, the paper develops a conceptual framework that situates large language model-based user modeling within the broader sociotechnical infrastructure of customer service. It considers how retrieval-augmented generation, memory mechanisms, and parameter-efficient adaptation can support continuous user understanding while limiting the risks of overpersonalization and opaque decision-making. The analysis highlights that effective query routing requires not only accurate intent recognition but also institutional alignment, escalation policies, and accountability mechanisms. The paper concludes by identifying future research directions for more transparent, resilient, and user-centered language model-based service platforms.

Keywords

large language models; user modeling; intelligent customer service; query routing; intent inference; sociotechnical systems.

1. Introduction

Customer service operations have shifted from scripted interactive voice response systems toward data-intensive conversational platforms that must interpret highly variable user

language. The development of bidirectional pretraining [1] and large-scale autoregressive language modeling [2] has substantially improved the ability of computational systems to represent linguistic nuance. These advances are not limited to text generation; they also enable new forms of user understanding because language models can infer preferences, constraints, affective signals, and situational context from unstructured utterances. When integrated into customer service, such models support a transition from static rule-based routing to context-aware query handling, in which user identity is continuously reconstructed from interaction evidence.

The architectural implications of this shift are significant. Instruction-tuned language models [3] and chain-of-thought prompting [4] allow systems to decompose ambiguous requests into interpretable subgoals before committing to a routing decision. The underlying transformer architecture [5] provides a flexible sequence encoder that can absorb multiple conversational turns, while unified text-to-text formulations [6] make it possible to treat intent classification, slot filling, summarization, response generation, and routing as related generative tasks. These modeling capabilities create the foundation for user models that are not static database records but probabilistic, contextual, and continuously revised representations.

This paper examines how large language model-based user modeling can be designed and governed within intelligent customer service systems. It focuses on system-level concerns rather than isolated model accuracy. The analysis addresses the integration of retrieval and memory, the inference of query intent, the routing of requests across human and automated channels, and the structural trade-offs that arise in large-scale deployment. The emphasis is on architecture, governance, robustness, fairness, and sustainability. By treating user modeling as an infrastructural practice rather than a single predictive component, the paper provides a forward-looking account of how language model systems can support responsible and effective customer service.

2. Background and Conceptual Foundations

Language models have evolved from task-specific architectures to general-purpose textual reasoning systems. Early work on unsupervised multitask learning [7] demonstrated that a language model trained on diverse web text could perform several tasks without task-specific supervision. This capability is important for customer service because user requests are heterogeneous in vocabulary, formality, and domain. At the same time, critical analyses of large-scale language modeling [8] have emphasized that fluency is not equivalent to reliable knowledge, and that generative systems may reproduce social biases or produce plausible but incorrect statements. A system-level perspective must therefore treat the language model as a source of probabilistic signal that requires verification, constraining, and interpretation.

The foundation model perspective [9] clarifies why language models can serve as reusable components across service functions. These models encode broad linguistic and world knowledge, but they require adaptation layers for organizational specifics. Retrieval-augmented generation [10] and dense passage retrieval [11] are particularly relevant for customer service because they allow user-facing responses and routing decisions to be grounded in policy documents, product databases, and interaction histories. Surveys of large language models [12] further document the rapid expansion of adaptation methods, evaluation protocols, and deployment strategies. Instruction scaling [13] has shown that models can be tuned to follow organizational procedures more reliably, which is essential when customer service systems must adhere to legal and policy constraints.

Classical user modeling research provides a useful conceptual anchor. Early stereotype-based models [14] reduced users to predefined categories, while generic user modeling systems [15] introduced more flexible representations that could be updated as new evidence arrived. Recommender system research [16] similarly moved from static preference vectors to contextual and sequential models. Large language models can be understood as extending this trajectory by allowing user models to be constructed directly from natural language evidence. A language-based user model can represent explicit requests, inferred goals, emotional tone, and conversational history in a shared representation space, thereby reducing the need for rigid feature engineering.

3. User Modeling Architecture for Customer Service

A central architectural challenge is how to build and maintain user models that are sufficiently rich for routing decisions without becoming computationally prohibitive. In many customer service systems, user identity is fragmented across account records, purchase histories, prior chat transcripts, and device information. Large language models can ingest this heterogeneous evidence as text and produce a coherent user state. Recent work on language model-based user memory systems [17] illustrates how search and matching can be improved by maintaining interpretable memory representations that support query understanding over time. Such approaches suggest that user modeling should be treated as a dynamic summarization process rather than a static lookup operation.

The design space can be organized around three interacting components: evidence encoding, user state aggregation, and downstream conditioning. Evidence encoding converts raw interaction data into textual descriptions that preserve relevant context. User state aggregation updates a working representation as new turns arrive, balancing recent information against long-term history. Downstream conditioning then passes this representation to routing, response generation, or escalation modules. Dialogue system research [18] has long recognized that conversation state and user state are distinct but coupled. In language model-based systems, the coupling is stronger because the same model may simultaneously infer user goals, track sentiment, and predict the most appropriate service channel.

A key structural trade-off concerns the placement of user modeling within the service pipeline. In tightly coupled designs, the user model is recomputed at every turn and directly influences all downstream decisions. This approach improves responsiveness but increases latency and can propagate errors. In loosely coupled designs, the user model is updated asynchronously and cached, reducing computational load but potentially missing rapid shifts in user intent. Hybrid configurations may use lightweight incremental updates for immediate routing and more expensive periodic re-computation for long-term profile refinement. The appropriate design depends on contact volume, session duration, risk tolerance, and the cost of misrouting.

Memory architecture is another important consideration. Short-term memory captures the current conversation, while long-term memory retains preferences, prior resolutions, and recurring issues. Language models can support both if integrated with retrieval mechanisms that select relevant historical evidence at inference time. However, memory retrieval introduces trade-offs between recall and precision. Selecting too much history may distract the model, while selecting too little may remove important context. System designers must therefore configure retrieval scope, recency weighting, and memory consolidation policies in ways that align with service objectives and privacy constraints.

4. Query Routing and Intent Inference

Query routing in intelligent customer service requires more than identifying a topic label. It involves estimating the user's desired outcome, the urgency of the issue, the appropriate resolution channel, and the likelihood that automated handling will be successful. Language model-based user modeling can support this by generating structured interpretations of user intent from conversational context. These interpretations may include product categories, issue types, emotional intensity, and preference for self-service or human assistance. Because the same language model can perform intent extraction and routing recommendation, the system can maintain consistency between user understanding and action selection.

Routing decisions are increasingly framed as sequential decision problems under uncertainty. A system must decide whether to answer directly, request clarification, escalate to a human agent, or transfer to a specialized team. Language models can generate clarifying questions when intent is ambiguous, reducing premature routing. They can also estimate confidence by examining the consistency of generated interpretations across different prompts or sampling strategies. Trust in such systems depends on the alignment between stated confidence and actual decision quality. Formal analyses of trust in artificial intelligence [19] emphasize that users and operators need evidence of competence, reliability, and explainability, which is especially relevant when routing decisions determine access to human support.

A major risk in language model-based routing is overconfidence in fluent but uncertain interpretations. A model may confidently assign a request to a billing queue when the user is actually reporting a technical fault, especially if the vocabulary overlaps. This risk is amplified in multilingual settings and in domains with rapidly changing products. System-level mitigation requires redundancy in intent signals, such as combining language model outputs with rule-based checks, retrieval similarity scores, and operational metadata. Escalation thresholds can be adjusted dynamically based on estimated user frustration, prior contact history, or the regulatory sensitivity of the issue. The routing architecture should thus be designed as an ensemble of complementary signals rather than a monolithic model decision.

The relationship between user modeling and routing also raises questions about user agency. If a user model predicts that a customer prefers self-service, the system may withhold visible pathways to human agents, creating a frustrating experience when the prediction is wrong. Conversely, always offering human escalation can overwhelm human capacity and reduce operational efficiency. A well-governed routing system should make its user model actionable but contestable. Customers should be able to correct misunderstandings, request different service channels, and understand why a particular route was selected. This requires the user model to expose interpretable evidence and for routing policies to include explicit override procedures.

5. System-Level Infrastructure and Deployment

Deploying large language model-based user modeling in customer service involves substantial infrastructural coordination. These systems require data pipelines that transform interaction logs into training and retrieval corpora, model serving infrastructure that can handle variable request volumes, and monitoring systems that detect drift, latency degradation, and policy violations. The hidden technical debt associated with machine learning systems [22] is particularly acute when language models are embedded in customer service because the boundaries between data, model, and policy are complex and constantly shifting. Small changes in product catalogs, refund policies, or escalation rules can invalidate assumptions encoded in the user model.

Parameter-efficient adaptation methods reduce the cost of customizing large language models for specific service domains. Techniques such as adapter-based tuning [23] and low-rank adaptation [24] allow organizations to update model behavior without retraining the full base model. This is important for customer service because policies are frequently revised, and different business units may require distinct user modeling behaviors. Tooling such as the Transformers library [25] provides standardized interfaces for loading, adapting, and serving these models. Nevertheless, standardization introduces its own governance challenges, because reliance on shared model artifacts can obscure the provenance of specific behaviors and complicate accountability.

A resilient deployment architecture separates the user modeling layer from the response generation and routing decision layers. This separation allows independent evaluation and rollback. For example, if a new user memory consolidation policy degrades routing accuracy, the change can be reverted without retraining the response generator. It also facilitates compliance with data retention policies, because user model components can be configured to expire or anonymize features according to jurisdictional requirements. In global customer service operations, data residency and cross-border transfer restrictions may require geographically distributed user model repositories, with language models performing inference at the edge or in regional data centers.

Monitoring and evaluation must extend beyond aggregate accuracy metrics. Operational dashboards should track session-level outcomes such as first-contact resolution, escalation rates, customer satisfaction, and average handling time, disaggregated by user segment and intent category. Language model-based user modeling can create feedback loops if routing decisions influence the distribution of future user interactions. For example, if a model routes most billing questions to self-service, fewer billing conversations may appear in the assisted channel, making the model appear more accurate than it is. Evaluation must therefore account for selection effects and use counterfactual or randomized audit mechanisms to estimate routing quality under alternative policies.

6. Robustness, Fairness, and Governance

Language model-based user models inherit the robustness limitations of their underlying models. They may fail under dialect variation, nonstandard syntax, code-switching, or domain-specific terminology that is underrepresented in pretraining. In customer service, such failures can produce systematic disparities in routing outcomes. Bias and fairness research [20] highlights that machine learning systems can discriminate through feature selection, label imbalance, and feedback loops. Language models add a further layer of risk because they encode social stereotypes and may associate certain language styles with lower urgency or competence. A user model that interprets nonnative phrasing as confusion may route multilingual customers away from human assistance, compounding service inequities.

Fairness in query routing requires careful definition and measurement. It is not sufficient to require equal routing rates across groups, because different groups may have different needs and preferences. More appropriate criteria include equal accuracy across groups, equal rates of appropriate human escalation for similar issue severity, and equal opportunity to access specialized support. Auditing these properties requires collecting sensitive attributes where legally permissible, or using proxy methods to estimate group differences without direct demographic identification. In many jurisdictions, privacy regulations restrict the use of demographic data, making fairness evaluation technically and legally complex. Governance

frameworks should therefore combine quantitative audits with qualitative evaluations of user experience and complaint patterns.

Ethical guidance for artificial intelligence [21] emphasizes principles such as transparency, accountability, privacy, and human oversight. These principles have concrete implications for language model-based user modeling. Transparency requires that users understand when automated systems are interpreting their language and making routing decisions. Accountability requires that organizations can trace a routing outcome to specific model versions, data sources, and policy configurations. Privacy requires that language model training and inference do not unnecessarily expose personal information. Human oversight requires that high-risk decisions, such as denial of service or legal escalation, can be reviewed by qualified personnel. These requirements are not purely technical; they depend on organizational roles, audit processes, and regulatory alignment.

Governance mechanisms should include model cards, data documentation, and routing policy reviews. A user modeling system should be treated as a living socio-technical artifact whose behavior emerges from training data, prompting strategies, retrieval corpora, and business rules. When a routing error causes harm, it is rarely attributable to a single component. Effective governance therefore requires cross-functional review boards that include data scientists, customer service managers, legal experts, and user advocates. The increasing use of language models in high-volume customer service makes such governance mechanisms operationally necessary rather than optional.

7. Sustainability and Long-Term Adaptation

Sustainability in language model-based customer service involves computational efficiency, financial cost, and organizational maintainability. Large language models consume significant energy during training and inference, and high-volume customer service can amplify operational costs. Techniques such as caching frequent user model embeddings, distilling smaller models for low-complexity routing, and using parameter-efficient adapters can reduce resource use. However, cost reduction should not compromise service quality for less common user groups or low-resource languages. Sustainability also includes the human labor required to annotate data, review escalations, and maintain policy knowledge. If a system constantly requires manual correction, its long-term value is questionable.

A sustainable user modeling architecture should support incremental learning without catastrophic forgetting of established policies. In customer service, organizations must preserve consistent treatment of recurring issues while adapting to new products and procedures. Retrieval-augmented memory can help by keeping policy knowledge external to the model weights, allowing updates without full retraining. User models can also be aligned with knowledge bases that are versioned and auditable. The system should distinguish between transient conversational state and durable user preferences, applying different update frequencies and retention rules to each.

Long-term adaptation requires evaluation across multiple timescales. Short-term metrics may reward systems that resolve conversations quickly, while long-term outcomes depend on customer retention, trust, and brand perception. A user model that aggressively steers users to self-service may improve quarterly operational metrics but erode customer loyalty if users feel unheard. Longitudinal studies and cohort analyses are necessary to understand how routing policies affect user behavior over time. The integration of language model-based user

modeling should therefore be accompanied by organizational learning processes that revisit assumptions and adjust priorities.

The environmental footprint of inference is also a governance concern. High-volume routing decisions can be supported by smaller specialized models or by selective invocation of large models only for ambiguous cases. Cascaded architectures can use lightweight intent classifiers for high-confidence queries and reserve full language model inference for complex or high-risk interactions. This approach aligns sustainability with robustness, because it reduces exposure to generative errors in routine cases while preserving sophisticated reasoning where it is most valuable. It also makes cost structures more predictable for service organizations operating under fixed budgets.

8. Future Research Directions

Several research directions merit further investigation. First, the development of user modeling benchmarks that reflect real customer service distributions is needed. Current natural language benchmarks often emphasize clean text and single-turn tasks, whereas customer service involves noisy input, long conversations, and shifting policies. Evaluation should measure not only intent accuracy but also the quality of routing decisions under uncertainty, the fairness of outcomes across user groups, and the interpretability of user models. Such benchmarks should be accompanied by operational datasets that preserve privacy while enabling reproducible research.

Second, methods for uncertainty-aware routing require further development. Language models can generate plausible interpretations, but they often lack calibrated confidence estimates. Research is needed on how to combine sampling-based uncertainty, retrieval confidence, and user feedback to decide when clarification or human escalation is appropriate. This is particularly important in regulated sectors where incorrect routing can delay access to critical services. Systems should be able to recognize gaps in their user models and explicitly request missing information rather than silently making assumptions.

Third, the governance of language model-based user memory systems requires more detailed study. Memory retention, deletion, and access control are central to user autonomy and regulatory compliance. Future work should develop technical mechanisms for selective forgetting, consent-aware memory construction, and user-facing explanations of how memory influences routing. The interaction between privacy regulations and language model memorization remains an open challenge, especially for systems that process sensitive financial, health, or legal information.

Fourth, cross-organizational and multi-agent customer service ecosystems introduce new coordination problems. Users may interact with multiple service providers that share partial user models or route requests across organizational boundaries. Language model-based systems could support negotiation among agents, but this raises questions about data sharing, liability, and consistency. Research on decentralized user modeling and federated learning may offer pathways for collaborative service improvement without centralizing sensitive data. Such approaches must be evaluated against the risks of model poisoning, inference attacks, and uneven participation.

Finally, human-AI collaboration in routing decisions deserves deeper attention. Human agents can provide corrective signals that improve user models over time, but only if their feedback is captured without imposing excessive annotation burdens. Designing interfaces that allow agents to indicate why a routing decision was incorrect can generate rich training data for

future model improvements. At the same time, organizations must avoid overburdening agents with model supervision tasks that distract from customer care. The integration of language model-based user modeling into customer service should ultimately serve both operational efficiency and the quality of human interaction.

9. Conclusion

Large language model-based user modeling offers a transformative opportunity for intelligent customer service and query routing by enabling richer, more adaptive, and more context-aware representations of users. However, realizing these benefits requires careful attention to system architecture, operational infrastructure, robustness, fairness, and governance. This paper has presented a system-level analysis of how language models can be integrated into user modeling pipelines, how they support intent inference and routing, and what structural trade-offs emerge in large-scale deployment. It has emphasized that user models should be dynamic, interpretable, and contestable, and that routing decisions must be embedded within accountable institutional processes.

The discussion has also highlighted the importance of memory systems, retrieval augmentation, and parameter-efficient adaptation in building sustainable customer service platforms. These techniques enable organizations to leverage the general capabilities of large language models while maintaining control over policy knowledge and operational cost. At the same time, they introduce new challenges related to privacy, bias, drift, and evaluation. Future research should address these challenges through interdisciplinary approaches that combine natural language processing, human-computer interaction, organizational science, and regulatory analysis.

Ultimately, the success of large language model-based user modeling in customer service will depend not only on technical accuracy but also on institutional trust. Systems that route users fairly, explain their decisions, respect privacy, and allow human intervention are more likely to produce sustainable value. The field is moving toward a view of user modeling as a continuous, socially embedded process rather than a static classification task. This shift calls for architectures that are modular, memory-aware, and governed by clear accountability structures. By adopting such a perspective, researchers and practitioners can build intelligent customer service systems that are both effective and responsible.

References

1. 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.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
4. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
7. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
9. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
11. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... 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. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J.-R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
13. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
14. Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3(4), 329–354.
15. Kobsa, A. (2001). Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1–2), 49–63.
16. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
17. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
18. Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2), 25–35.
19. Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635.
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

22. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
23. Houlisby, N., Giurciu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., ... Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*, 2790–2799.
24. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations*.
25. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.