International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

# Privacy-Preserving Large Language Model-Based Recommendation Systems - Challenges, Techniques, and Opportunities

**Ying Li**[1]

[1]*Meta Platform Inc., USA*

---

**Abstract -** *As large language models (LLMs) increasingly serve as the backbone for intelligent recommendation systems, they offer unprecedented personalization capabilities through deep contextual understanding and natural language generation. However, this advancement also raises critical privacy risks, including memorization of sensitive data, prompt injection attacks, and cross-domain inference vulnerabilities. In this survey, we present a comprehensive review of privacy-preserving techniques for LLM-based recommendation systems. We outline various LLM-powered architectures, categorize the emerging threat landscape, and analyze mitigation strategies including multi-agent designs, federated learning, differential privacy, and pseudonymization. We also identify ongoing challenges such as limited domain-specific benchmarks, inadequate user control, and fairness-privacy trade-offs. Finally, we highlight future research directions, including adaptive privacy interfaces, explainability under privacy constraints, and unified frameworks for legal compliance and ethical alignment. Our goal is to establish a foundational understanding of privacy-preserving mechanisms in the context of LLM-driven recommendation systems and to inspire further innovations that balance personalization with responsible data stewardship.*

***Key Words***: Large Language Models, Recommendation Systems, Privacy-Preserving AI, Differential Privacy, Federated Learning, Contextual Integrity, Explainability

## 1.INTRODUCTION

The integration of large language models (LLMs) into recommender systems has significantly transformed how personalized digital experiences are delivered. These models enhance the quality of recommendations by enabling more sophisticated natural language understanding, richer contextual reasoning, and more interactive user engagement. However, their adoption has also introduced a range of new privacy concerns that traditional recommendation frameworks were not designed to handle. This survey provides a comprehensive exploration of the current landscape of privacy-preserving techniques within LLM-powered recommendation systems. We outline a new taxonomy of privacy risks, examine the mitigation strategies being developed to address them, and analyze the architectural patterns emerging in this space. The goal is to offer a clear roadmap for future research at the convergence of LLMs, personalization, and privacy.

## 2. LLM-POWERED RECOMMENDER ARCHITECTURES

### 2.1 LLM as a Re-Ranker or Explainer

LLMs are now frequently used to refine the outputs of traditional recommendation engines by re-ranking results based on natural language input or by providing detailed explanations for recommendations. For instance, a collaborative filtering algorithm may initially generate a list of product suggestions, but an LLM can refine this list using a nuanced understanding of a query like "comfortable shoes for all-day city walking." In addition to enhancing relevance, LLMs can generate natural-sounding explanations—e.g., "This book matches your interest in character-driven mysteries"—which builds user trust and improves system transparency without changing the core algorithmic mechanisms.

### 2.2 LLM as End-to-End Recommender

This architecture treats the recommendation task as one of text generation. Rather than selecting from a fixed set of options, the system generates recommendations directly in natural language. For example, a travel platform might ask an LLM to propose a complete itinerary based on a user's preferences and constraints. These systems typically use prompt engineering and in-context learning to tailor results, which offers much more flexibility in capturing complex or evolving user preferences than traditional embedding-based methods.

### 2.3 LLM for User Preference Modeling

LLMs can dynamically construct user profiles by interpreting natural language inputs such as conversation snippets, reviews, or queries. Unlike traditional systems that depend on structured logs (e.g., clicks or purchase history), LLMs can infer preferences from casual, unstructured language. For example, an offhand remark about dietary restrictions might be detected and incorporated into food recommendations. Additionally, LLMs show promise in transferring knowledge across domains—for example, learning a user's aesthetic taste in clothing and applying it to home decor suggestions.

### 2.4 Agentic and Multi-Modal Systems

Some of the most forward-looking systems rely on LLMs functioning as agents that actively engage with users and external systems. These systems are capable of initiating

dialogue to clarify user needs, accessing third-party databases, processing multiple data formats (e.g., text, image, audio), and generating well-justified recommendations. A shopping assistant, for example, could gather input across several channels—text conversation, visual preferences, search queries—and synthesize these into a curated list of products. This architecture supports long-term personalization by maintaining state across sessions and adapting to evolving preferences.

## 3. PRIVACY THREATS IN LLM-BASED RECOMMENDATION SYSTEMS

### 3.1 Memorization and Data Leakage

Because LLMs are trained on vast datasets, including user interactions, they can inadvertently memorize and later reveal sensitive data. This poses significant risks in personalized recommendation scenarios. A model fine-tuned on user-specific data may, for instance, surface unusually precise suggestions that indirectly disclose private medical, financial, or behavioral details. Unlike traditional recommendation engines, which typically rely on anonymized feature vectors, LLMs may reproduce raw text or personal identifiers from training data—underscoring the need for robust privacy controls [1, 2].

### 3.2 Prompt Injection and Jailbreaking

Prompt injection attacks manipulate LLMs by embedding adversarial instructions within seemingly benign user inputs. These attacks can exploit a model's language-following behavior to bypass privacy filters or extract unauthorized data. In recommendation systems, this may lead to exposure of other users' preferences or profiles. As LLMs become more capable and more responsive to open-ended queries, securing them against these forms of manipulation is an urgent challenge [3, 4].

### 3.3 Indirect Profiling via Output Analysis

Even when no direct identifiers are shared, the patterns in recommendation outputs can still be revealing. For example, a user might be consistently recommended content related to a rare illness, which could imply their health condition. These risks arise from statistical correlations that are difficult to detect and often overlooked by traditional privacy-preserving techniques. Indirect profiling is particularly problematic in systems with long-term personalization, where behavioral patterns accumulate over time [5].

### 3.4 Cross-User and Cross-App Inference

In ecosystems where LLMs operate across different applications or serve multiple users, privacy leakage can occur through unintended cross-context information sharing. For example, insights gleaned from a user's behavior in a health app could subtly influence recommendations in an e-commerce platform. These issues are exacerbated by shared model infrastructure and multi-tenant environments, making strict privacy boundaries difficult to enforce [2, 6].

## 4. MITIGATION TECHNIQUES

**Table -1:** Mitigation Strategies for Privacy Threat

| Privacy Threat | Mitigation Strategies |
|---|---|
| Memorization and Data Leakage | Differential Privacy [10], Federated Learning [11] |
| Prompt Injection and Jailbreaking | Prompt Sanitization, Multi-Agent Filtering [7] |
| Indirect Profiling via Output Analysis | Semantically-Aware Pseudonymization [8], Local Filters [9] |
| Cross-User and Cross-App Inference | Model Segregation, Coordination Protocols [6] |

### 4.1 Multi-Agent Privacy Architectures

One effective approach to reducing privacy risk is to divide the recommendation process into discrete stages handled by separate agents. For instance, a detection agent might identify sensitive inputs, a sanitization agent could redact or transform them, and a generation agent would then use the filtered data to produce recommendations. This separation of duties limits the scope of exposure for each component and supports modular auditing and enforcement of privacy guarantees [7].

### 4.2 Semantically-Aware Pseudonymization

Instead of simply redacting sensitive entities, advanced pseudonymization methods replace them with contextually similar alternatives. This allows systems to maintain meaningful responses while obscuring personally identifiable details. For example, a user's reference to a specific health condition could be replaced with a more general category, preserving the utility of the interaction while improving privacy protection [8].

### 4.3 Local Privacy Filters with Contextual Awareness

Edge-deployed models can act as real-time privacy filters, preprocessing user inputs before transmitting them to cloud-based recommenders. These models assess the sensitivity of data based on context—for instance, recognizing that location data may be relevant in a fitness app during a workout session, but not in a general query. Context-aware filtering ensures that privacy is enforced dynamically and appropriately [9].

### 4.4 Differential Privacy and Federated Approaches

Differential privacy techniques inject statistical noise into outputs to protect individual data points, while federated learning enables collaborative model training without sharing raw data. These approaches are particularly well-suited for recommendation settings involving sensitive personal data, such as financial transactions or health histories [10, 11].

### 4.5 Neuron-Level Decoupling for Ethical Alignment

In deep neural networks, certain internal representations may encode multiple ethical dimensions simultaneously, such as privacy and fairness. Recent work suggests that identifying and suppressing these shared neurons allows the system to treat each dimension independently, reducing trade-offs and improving alignment with user values and regulatory expectations [12].

## 5. OPEN CHALLENGES

### 5.1 Lack of Domain-Specific Privacy Benchmarks

While several privacy benchmarks exist for evaluating general-purpose language models, they often fall short when applied to recommendation systems. These systems present unique challenges that aren't captured in traditional benchmarks—for example, the risk of sensitive data being revealed through recommendation patterns or cross-service data flows. A health-focused recommender system, for instance, might inadvertently leak past diagnoses through its medication suggestions. This highlights the pressing need for benchmarks that are specifically designed to assess privacy risks within different domains of recommendation, particularly those that handle highly sensitive data like healthcare, finance, or education [13].

### 5.2 Granular Context Modeling

Privacy is inherently contextual, yet current systems often struggle to detect and adapt to the changing nature of context during user interaction. A single user may shift between professional and personal modes throughout the day or use the same device for both sensitive and casual inquiries. These shifts in context are subtle and not easily detectable by most systems, resulting in privacy violations when systems fail to adjust their data handling appropriately. Adapting recommendation models to respect these dynamic privacy norms requires a more nuanced understanding of user intent and environmental cues [14].

### 5.3 Utility-Privacy Trade-offs

A recurring challenge in privacy-preserving systems is the delicate balance between maintaining recommendation utility and enforcing meaningful privacy protections. Overzealous privacy mechanisms can degrade personalization, leading to irrelevant suggestions that frustrate users and reduce engagement. Conversely, minimal privacy controls may expose users to serious risks. Achieving the right balance requires mechanisms that can adapt to varying levels of sensitivity based on context, domain, or even individual user preferences. Developing systems that dynamically calibrate this trade-off remains an open research area [15].

### 5.4 Limited User Control and Transparency

Most current recommender systems offer users very limited control over how their data is used—typically limited to binary choices such as opt-in or opt-out. This simplistic model fails to account for the wide range of privacy preferences users may have. For example, someone might be comfortable sharing workout data for fitness suggestions but not for marketing purposes. Implementing fine-grained, intuitive privacy controls that are also technically enforceable is a complex but necessary challenge. Moreover, these controls must be transparent, allowing users to understand and trust how their data is being used [16].

### 5.5 Multi-Agent Coordination

As recommender systems grow more complex and incorporate multiple components—such as edge devices, cloud servers, and third-party APIs—ensuring consistent privacy enforcement across these interconnected agents becomes increasingly difficult. Each component may have different access levels, privacy protocols, and even regulatory constraints, especially in cross-border scenarios. Without robust coordination mechanisms, data may slip through the cracks, violating user expectations or legal requirements. Addressing this challenge calls for both technical solutions and improved governance models that can align the behavior of distributed systems [6].

## 6. FUTURE OPPORTUNITIES

### 6.1 Adaptive Privacy Interfaces

There is a growing interest in developing natural language interfaces that allow users to define their privacy preferences in conversational terms. Imagine a user saying, “Don’t use my shopping history when making health recommendations.” A well-designed system would not only understand this intent but enforce it accurately and consistently across all interactions. These interfaces lower the barrier to privacy customization, especially for non-technical users, and pave the way for a more human-centered approach to privacy in recommendation systems [17].

### 6.2 Lightweight Privacy Agents

With advances in model compression and on-device AI, it is now feasible to deploy compact, privacy-aware agents directly on user devices. These agents can intercept and filter

sensitive data before it ever reaches the cloud, providing real-time, context-sensitive privacy protection. For example, a smartphone-based agent might remove sensitive phrases from a query before passing it to a recommendation engine. Such systems offer a promising alternative to centralized privacy controls, shifting enforcement closer to the user and enhancing both privacy and trust [18].

### 6.3 Unified Fairness-Privacy Optimization

Privacy and fairness are often treated as separate design objectives, but recent research has shown that optimizing for one can unintentionally undermine the other. For instance, privacy-preserving techniques that obscure user demographics may reduce the accuracy of recommendations for underrepresented groups. Joint optimization frameworks that balance both objectives—perhaps through multi-objective training or architectural decoupling—can help ensure that systems remain inclusive while still protecting user data [19].

### 6.4 Legal-Compliant Benchmarks

Regulatory frameworks like the GDPR and CCPA provide valuable guidance on what privacy means in a legal context, but they often lack concrete technical interpretations. Future benchmarks could bridge this gap by evaluating systems based on their ability to comply with legal principles such as data minimization, right to object, and purpose limitation. Such benchmarks would not only guide the development of privacy-aware systems but also help organizations demonstrate regulatory compliance in measurable terms [13].

### 6.5 Privacy-Preserving Explainability

Explainability is a core component of user trust, but it can conflict with privacy when explanations inadvertently disclose sensitive reasoning paths. For example, telling a user that a film was recommended because they watched a highly personal or stigmatized title may be informative but also invasive. The next wave of explainable recommendation systems must find ways to provide useful insights without compromising user privacy. This might involve generating abstract explanations or using differential privacy techniques to mask sensitive features [20].

## 7. CONCLUSIONS

LLM-powered recommendation systems hold immense potential to transform how individuals discover content, services, and opportunities. At the same time, they introduce new and complex privacy risks that must be carefully addressed to ensure ethical and sustainable deployment. This survey has mapped the emerging privacy landscape in this domain, highlighting key risks such as memorization, prompt injection, and cross-context inference, as well as mitigation strategies like pseudonymization, differential privacy, and multi-agent architectures. We have also outlined ongoing challenges and promising future directions—from user-facing privacy controls to legal-compliant benchmarks and explainable, privacy-aware recommendations. As these systems become increasingly integral to everyday digital life, building robust, human-aligned privacy frameworks will be essential to their long-term success.

## REFERENCES

[1] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Erlingsson, Ú. (2021). Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium.

[2] Nasr, M., Shokri, R., & Houmansadr, A. (2023). Comprehensive privacy analysis of training data in machine learning models. IEEE Transactions on Information Forensics and Security.

[3] Zhou, S., Wang, Y., & Liu, X. (2023). Prompt injection attacks against language models. In Findings of the Association for Computational Linguistics.

[4] Perez, E., Kiela, D., & Cho, K. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

[5] Mireshghallah, F., Tramer, F., & Shokri, R. (2023). Privacy risks in language model outputs. In Proceedings of the IEEE Symposium on Security and Privacy.

[6] Luo, W., Zhang, Q., & Liu, Y. (2024). Cross-context privacy risks in multi-tenant AI systems. In Proceedings of the AAAI Conference on Artificial Intelligence.

[7] Guo, A., Sharma, S., & Wang, Y. (2024). Enhancing contextual privacy via multi-agent reasoning. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.

[8] Dou, Z., Liang, Y., & Li, X. (2023). LOPSIDED: Semantically-aware pseudonymization for language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

[9] Fan, J., Liu, H., & Raji, I. D. (2024). Protecting users from themselves: Contextual pre-filtering for privacy in LLMs. In Proceedings of the 2024 Conference on Human Factors in Computing Systems (CHI).

[10] Dwork, C. (2006). Differential privacy. In International Colloquium on Automata, Languages, and Programming (pp. 1-12). Springer.

[11] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private language models

without central coordination. arXiv preprint arXiv:1808.00510.

[12] Elhage, N., Nanda, N., Olsson, C., & Henighan, T. (2022). Toy models of superposition in machine learning. In Transformer Circuits Interpretability Papers.

[13] Li, T., Zhang, Y., & Shvartzshnaider, Y. (2024). PrivaCI-Bench: Evaluating privacy with contextual integrity. In Proceedings of the ACM Conference on Computer and Communications Security.

[14] Nissenbaum, H. (2010). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

[15] Huang, Z., Singh, A., & Sun, Y. (2023). Exploring the utility-privacy trade-off in generative recommender systems. In NeurIPS Workshop on Responsible Recommendation.

[16] Hardinges, J., Wang, R., & Mahdavi, M. (2024). Behavioral privacy and user agency in AI-powered personalization. In Proceedings of the ACM Conference on Human Factors in Computing Systems.

[17] Shvartzshnaider, Y., & Duddu, S. (2025). Survey of contextual integrity in language models. To appear in ACM Computing Surveys.

[18] Kumar, P., Tan, W., & Zhao, L. (2024). Safeguarding contextual privacy in prompt sharing. In Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning.

[19] Sun, L., Wang, J., & Lim, B. (2024). SPIN: Decoupling fairness and privacy in neural recommender systems. In Proceedings of the International Conference on Machine Learning.

[20] Mireshghallah, F., Tramer, F., & Shokri, R. (2024). Benchmarking contextual privacy in LLM explainability. In Proceedings of the 2024 Conference on Fairness, Accountability, and Transparency.