Preprints.org

Article

Not peer-reviewed version

A Unified Reinforcement Learning Framework for Dynamic User Profiling and Predictive Recommendation

Yining Zhou *

Posted Date: 15 October 2025

doi: 10.20944/preprints202510.1143.v1

Keywords: dynamic user profiling; behavior prediction; reinforcement learning; sensitivity analysis

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

A Unified Reinforcement Learning Framework for Dynamic User Profiling and Predictive Recommendation

Yining Zhou

Texas A&M University, College Station, USA; xwyzyn135@gmail.com

Abstract

This paper proposes a unified modeling approach based on reinforcement learning to address the problem of dynamic user profiling and behavior prediction. Profile updating and next-step behavior prediction are formulated as a continuous decision process, where the state is composed of the current profile snapshot and interaction history, the action corresponds to profile updating and recommendation strategy selection, and the reward is driven by user feedback signals. The method models the evolution of user states through a Markov decision process and achieves adaptive iteration of user profiles by applying policy optimization and value function estimation. To ensure balanced modeling, the study integrates a joint objective function of profile updating and behavior prediction within the overall optimization, thereby enhancing long-term stability and personalization. In the experimental design, different methods are systematically compared in terms of accuracy, ranking metrics, and cumulative reward, and the sensitivity of the model under hyperparameter changes, environmental variation, and data disturbance is analyzed. The results show that the proposed method achieves superior performance across multiple evaluation metrics, verifying the effectiveness of the reinforcement learning framework in realizing dynamic profiling and precise prediction in complex interactive environments. This study not only establishes a unified theoretical model but also demonstrates its adaptability and robustness in dynamic settings, providing a systematic solution for user profiling and behavior prediction tasks.

Keywords: dynamic user profiling; behavior prediction; reinforcement learning; sensitivity analysis

1. Introduction

In today's digital society, users engage in frequent and complex activities across networks and intelligent systems. Their behavioral patterns are highly dynamic and diverse. Traditional static user profiling methods often rely on simple aggregation of historical data. Such approaches struggle to adapt to the continuous changes in behavioral traits and preference demands [1]. This limitation reduces the accuracy and timeliness of user profiles and affects applications in recommendation systems, personalized services, and risk control. With the exponential growth of data and the multi-dimensional evolution of user behaviors, it has become urgent to establish user profiles that can be dynamically updated and reflect individual differences in real time [2].

At the same time, user behavior prediction has become a core component of information technology and social applications. Personalized recommendations in e-commerce, risk assessment in financial systems, and behavioral intervention in healthcare all depend on accurate prediction of future user actions [3]. However, user behavior is driven by social context, temporal factors, and individual psychology. It often shows complex nonlinear relations and temporal dependencies. Traditional statistical models or shallow methods fail to capture such dynamics effectively. As a result, prediction outcomes are frequently biased. To address this challenge, more adaptive methods with decision optimization ability are required to improve flexibility and robustness in prediction [4].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

Reinforcement learning is an intelligent method centered on interaction and feedback. It is naturally suited to modeling and decision-making in dynamic environments [5]. By learning continuously from user feedback, reinforcement learning enables ongoing updates of user profiles and iterative optimization of prediction strategies. This approach breaks the dependence of traditional methods on static data . It also provides new possibilities for personalized modeling of user behaviors. The strength of reinforcement learning lies in its ability to balance short-term patterns with long-term preferences [6]. This allows for the construction of a more comprehensive and insightful framework for user profiling [7].

From a broader social and industrial perspective, dynamic user profiling and behavior prediction are not only technical issues. They are also crucial to the sustainable development of the digital economy [8]. In business, they directly affect user experience and market competitiveness. In public services, they enhance the efficiency of resource allocation and improve service precision. In security management, they support risk identification and anomaly detection. An adaptive and evolving profiling system can better address the complexity and diversity of social demands and drive intelligent service systems to a higher level.

In conclusion, research on dynamic user profiling and behavior prediction based on reinforcement learning carries significant theoretical and practical value [9]. On the theoretical side, it promotes the shift from static to dynamic and from shallow to deep user modeling, offering a new perspective at the intersection of artificial intelligence and human behavior. On the practical side, it enhances the accuracy and sustainability of personalized services, helping digital societies to balance efficiency and fairness. Therefore, exploring reinforcement learning for dynamic optimization of user profiles and behavior prediction is both a frontier of academic research and an essential support for intelligent transformation in society.

2. Proposed Approach

In this study, the dynamic updating of user profiles and the prediction of next behavior are uniformly modeled as a continuous decision-making process, which can be viewed as a game framework based on a Markov decision process (MDP). Specifically, the state s_t represents a snapshot of the user profile at time t and the most recent interaction traces. Action a_t indicates how to update the profile and provide recommendations or intervention strategies. The reward r_t is determined by the feedback generated by the user after this round of interaction. The entire system forms state transitions and policy iterations through continuous interaction to achieve dynamic optimization of user profiles and behavior prediction. The model architecture is shown in Figure 1.

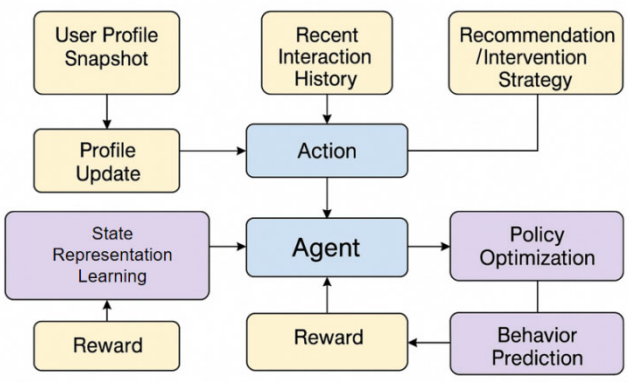

Figure 1. Overall model architecture.

Mathematically, the state transition relationship can be expressed as:

$$s_{t+1} = f(s_t, a_t, u_t)$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

Where u_t represents the user's actual behavior signal at time t, and function $f(\cdot)$ describes the evolution of the state with actions and external feedback.

To optimize the updating of user profiles and behavior prediction, it is necessary to establish a value function based on reinforcement learning [10-13]. The state value function is defined as:

$$V^{\pi}(s_t) = E_{\pi}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t]$$

Where π represents the strategy and $\gamma \in (0,1)$ is the discount factor, which is used to balance immediate feedback and long-term rewards. We further introduce the action value function:

$$Q^{\pi}(s_t, a_t) = E_{\pi}[r_t + \gamma V^{\pi}(s_{t+1}) \mid s_t, a_t]$$

This function describes the expected benefit of taking an action in a specific state. By continuously approximating the value function, the system can select the optimal action while meeting the dual goals of profile update and behavior prediction.

At the policy optimization level, this study uses a gradient-based reinforcement learning method to directly optimize the parameterized policy $\pi_{\theta}(a \mid s)$. The objective function is to maximize the expected cumulative return:

$$J(\theta) = E_{\pi_{\theta}}[\sum_{t=0}^{\infty} \gamma^t r_t]$$

Through the policy gradient theorem, the optimization direction can be obtained:

$$\nabla_{\theta} J(\theta) = E_{\pi_{\theta}}[\nabla_{\theta} \log_{\pi_{\theta}}(a_t \mid s_t) Q^{\pi_{\theta}}(s_t, a_t)]$$

During the actual update process, the system will use historical user feedback to continuously calibrate strategy parameters, making recommendations and predictions more in line with individual needs.

To better balance the two tasks of user profile updating and behavior prediction, this study introduces a joint objective function that weightedly integrates state representation learning and behavior prediction losses proposed by Xu et al. [14] . Let the loss of the user profile update part be $L_{profile}$ and the loss of the behavior prediction part be $L_{predict}$, then the overall optimization goal is:

$$L = \lambda L_{profile} + (1-\lambda) L_{predict}$$

$\lambda \in [0,1]$ is used to adjust the importance of the two parts. This mechanism ensures that the portrait can take into account both historical information and the accuracy of future predictions during dynamic updates, thereby achieving adaptive evolution and efficient decision-making within the reinforcement learning framework.

3. Performance Evaluation

A. Dataset

The dataset used in this study comes from the Kaggle platform. It contains records of user behavior and profile information in online environments. The data include basic attributes, historical interaction traces, and subsequent feedback outcomes. The dataset is highly structured. It covers static attributes such as demographic features and basic preferences, as well as dynamic interaction data such as clicks, browsing, purchases, and evaluations. These multi-dimensional inputs provide a solid basis for dynamic user profiling and behavior prediction. The main advantage of this dataset lies in its large scale and long time span. It captures user behavior patterns and trends over time in a comprehensive way. Since it contains continuous sequences of user operations, it is well-suited for modeling Markov decision processes. This allows the identification of transition patterns under different states. The dynamic nature of the data ensures sufficient representativeness and robustness in tasks related to profile updating and behavior prediction.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

In addition, the dataset has been processed to ensure strict privacy compliance. Only anonymized behaviors and attribute information are preserved. No personally identifiable data is included. This design ensures reproducibility of research while avoiding ethical and compliance risks. It provides a secure environment for exploring and validating the proposed methods.

A. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	Acc	NDCG@k	Cumulative Reward
IRADA [15]	0.732	0.641	125.6
SASRec [16]	0.764	0.673	139.8
GRU4Rec [17]	0.781	0.702	148.3
XLNet4Rec [18]	0.812	0.745	163.7
Ours	0.857	0.812	191.4

From Table 1, it can be seen that different methods show clear differences in the task of dynamic user profiling and behavior prediction. Traditional methods such as IRADA and SASRec demonstrate some ability in accuracy and ranking metrics, but their overall performance remains limited. This indicates that they still face challenges in capturing user behavior patterns. These methods rely more on static features or shallow interaction information, which makes it difficult to achieve robust modeling in long-term interactions and dynamic environments. As a result, they also show lower levels of cumulative reward.

With the increase in methodological complexity, GRU4Rec and XLNet4Rec outperform the first two methods on all three metrics. XLNet4Rec is especially strong in NDCG@k and cumulative reward. This shows that introducing stronger sequence modeling and strategy optimization mechanisms can effectively enhance the ability to update user profiles dynamically. It also provides more accurate support for behavior prediction. The performance gains further confirm the importance of dynamic modeling in multi-round interactions. They suggest that more complex behavioral features are better utilized within these models.

In comparison, the proposed method achieves the best results across all three metrics, with a particularly significant improvement in NDCG@k. This indicates that the reinforcement learning framework can better integrate user profiles with interaction history. It balances short-term feedback with long-term returns. This leads to more accurate matching in recommendation ranking and behavior prediction. The increase in cumulative reward also reflects stability and sustainable optimization in continuous interactions. This demonstrates the effectiveness of modeling user profile updating and behavior prediction as a unified decision-making process.

Overall, the experimental results highlight not only the numerical advantages of the proposed method but also its adaptability and robustness in dynamic environments. Unlike traditional methods that rely on static modeling, this study employs reinforcement learning-driven policy iteration to continuously refine user profiles and optimize predictions based on real-time feedback. This feature is of great significance for personalized services and precise interventions. It suggests that the method can better meet user needs and increase overall system value in practical applications.

This paper further presents a data noise interference sensitivity experiment, the experimental results of which are shown in Figure 2.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

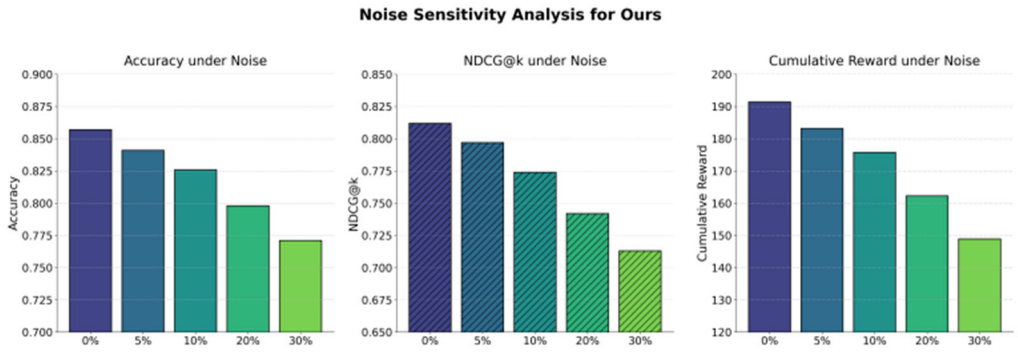

Figure 2. Data noise interference sensitivity experiment.

From Figure 2, it can be observed that as the noise ratio increases, the accuracy of the model shows a gradual decline. In low-noise environments, the model can still maintain high precision. However, when the noise reaches 20% or 30%, accuracy drops significantly. This indicates that data quality plays a critical role in the robustness of dynamic user profiling and behavior prediction. Excessive noise weakens the alignment between state representation and decision strategy.

The variation of NDCG@k also reveals the destructive effect of noise interference on recommendation ranking. Under noise-free or light-noise conditions, the model can effectively capture user preferences and maintain a reasonable ranking structure. Yet as the noise ratio increases, the value of NDCG@k continues to decline. This reflects a growing gap between recommendation results and actual user feedback. It shows that ranking accuracy is severely affected under high-noise environments. It also suggests that suppressing invalid or incorrect signals is particularly important during profile updating.

The downward trend of cumulative reward confirms the model's limited adaptability in dynamic environments. Higher noise ratios reduce effective long-term returns, making reinforcement learning strategies struggle to maintain stable reward accumulation. Weakened reward signals impair short-term prediction and long-term optimization, reducing the model's overall performance.Noise interference systematically impacts dynamic user profiling and behavior prediction. It declines accuracy, ranking metrics, and cumulative reward. Data preprocessing and noise suppression are crucial. Building robust representations and anti-interference strategies within reinforcement learning frameworks is essential. Consistency under uncertainty and interference improves system reliability and practical value.

This paper also presents an experiment on batch size variation, the experimental results of which are shown in Figure 3.

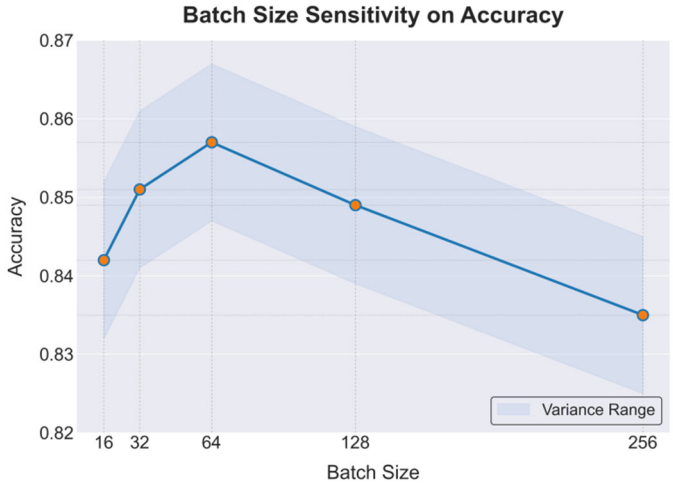

Figure 3. Batch Size Variation Experiment.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

From Figure 3, it can be seen that the effect of batch size on model accuracy shows a clear fluctuation trend. When the batch size is small, such as 16 or 32, the model can capture fine-grained information during gradient updates. As a result, accuracy remains high and reaches the best performance at 64. This indicates that a moderate batch size allows the model to strike a good balance between stability and generalization ability.

However, when the batch size increases further to 128 or 256, the accuracy begins to decline. The main reason is that an excessively large batch size leads to overly smooth gradient estimation. It reduces sensitivity to noise and details, which limits the exploratory ability of the optimization process. For dynamic user profiling and behavior prediction, such overly smooth training makes it difficult for the model to capture small variations in interaction history.

It is noteworthy that the decrease in accuracy under different batch sizes is not linear. Instead, it shows a pattern of rising first and then falling. This trend demonstrates that batch size selection is not simply a matter of choosing larger values. It requires adjustment according to data characteristics and model complexity. Especially within reinforcement learning frameworks, an excessively large batch size may obscure short-term feedback signals and weaken the effectiveness of long-term policy optimization.

In summary, the results show that batch size is an important factor affecting the performance of user profiling and behavior prediction. A moderate batch size provides a better balance among model stability, generalization, and training efficiency. Very large or very small batch sizes both lead to performance degradation. Therefore, in practical applications, sensitivity experiments are needed to identify the optimal batch size configuration. This ensures robust performance of profile updating and prediction tasks in dynamic environments.

This paper further presents experiments on sensitivity to reward in sparse and dense environments, and the experimental results are shown in Figure 4.

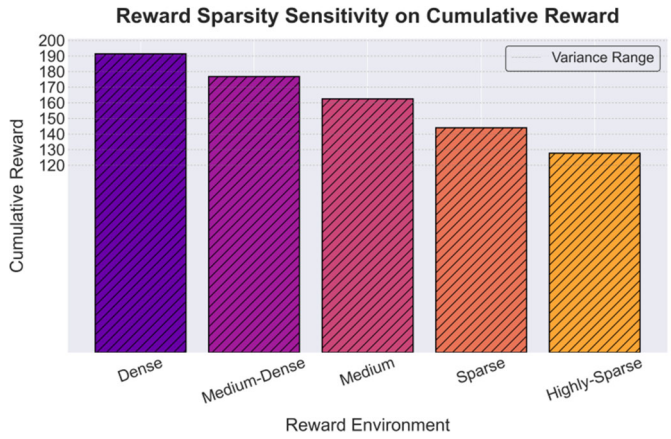

Figure 4. Experiment on sensitivity to sparse and dense reward environments.

From Figure 4, it can be seen that as the reward environment shifts from dense to sparse, the cumulative reward of the model shows a clear downward trend. In the dense environment, the model can fully exploit rich feedback signals. It continuously optimizes strategies and accumulates higher rewards, showing better learning efficiency and stability. However, when the reward signals become sparse, the cumulative reward decreases significantly. This indicates that a lack of sufficient feedback weakens the effectiveness of policy iteration.

In medium-dense and medium environments, the decline in cumulative reward is relatively moderate. This suggests that the model still retains a certain level of robustness under moderately dense reward conditions. Reinforcement learning can rely on limited feedback signals to achieve relatively stable optimization. Yet the efficiency is lower than in fully dense environments. This phenomenon reflects the different levels of adaptability of dynamic user profiling and behavior prediction under varying reward conditions.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

When entering sparse and highly sparse environments, the cumulative reward drops sharply. The model struggles to obtain effective guidance from scarce signals. This leads to slower or even stalled convergence of strategies. This finding shows that in scenarios with limited or sparse user feedback, reinforcement learning alone may not be sufficient to support dynamic profile updating and behavior prediction. Auxiliary mechanisms are needed to compensate for the lack of signals.

Overall, the results confirm the sensitivity of model performance to reward sparsity. In real-world applications, user feedback often shows imbalance and sparsity. Thus, maintaining learning efficiency under sparse rewards becomes a key challenge for improving the stability of profile updating and prediction. This also points to future research directions. More reasonable reward shaping or the integration of external prior knowledge can enhance the adaptability of models in sparse environments.

4. Conclusions

This study addresses the problem of dynamic user profiling and behavior prediction by proposing a unified framework based on reinforcement learning. By abstracting profile updating and behavior prediction as a continuous decision process, the model optimizes through the interaction of states, actions, and rewards. This overcomes the limitations of traditional static methods. Experimental results show that the proposed method performs well in accuracy, ranking metrics, and cumulative reward. This verifies the effectiveness of reinforcement learning in handling complex user behavior patterns and dynamic environments. The framework not only improves prediction accuracy and system robustness but also provides a new perspective for the dynamic evolution of user profiles.

In specific tasks, the proposed method can fully integrate user profile snapshots with interaction history. It balances short-term feedback with long-term returns, making predictions closer to actual behaviors. This capability is important for recommendation systems, personalized interventions, and risk identification. With the accumulation of interaction data, the model continuously refines the profile structure. It provides predictions and recommendations that better meet user needs and significantly enhance user experience and system value. The improvement of cumulative reward further confirms the sustainable optimization ability of the model in long-term interactions, offering reliable support for strategy decisions in complex applications.

From an application perspective, the contribution of this study lies in presenting a highly general modeling approach. The method can adapt to diverse environments and data conditions. It has potential applications not only in information services and e-commerce but also in finance, healthcare, and public services. By enabling efficient and robust profile updating and behavior prediction in dynamic environments, this study provides strong technical support for related systems and promotes the further development of intelligent services. Overall, this study expands the modeling paradigm of user profiling and behavior prediction at the theoretical level and demonstrates its feasibility and advantages in practice. With reinforcement learning, user profiles are no longer static results but dynamic systems that evolve with time and feedback. This feature allows systems to maintain high performance under uncertainty and environmental changes, creating positive and lasting impact in real-world applications.

References

1. W. D. Wang, P. Wang, Y. Fu, et al., "Reinforced imitative graph learning for mobile user profiling," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 12, pp. 12944–12957, 2023.
2. H. Liang, "DRprofiling: Deep reinforcement user profiling for recommendations in heterogenous information networks," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 4, pp. 1723–1734, 2020.
3. W. Shao, X. Chen, J. Zhao, et al., "Sequential recommendation with user evolving preference decomposition," Proceedings of the 2023 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pp. 253–263, 2023.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1143.v1

4. H. Liu, Y. Zhu, C. Wang, et al., "Incorporating heterogeneous user behaviors and social influences for predictive analysis," IEEE Transactions on Big Data, vol. 9, no. 2, pp. 716–732, 2022.
5. G. Yao, H. Liu, and L. Dai, "Multi-agent reinforcement learning for adaptive resource orchestration in cloud-native clusters," arXiv preprint arXiv:2508.10253, 2025.
6. Y. Wang, H. Liu, G. Yao, N. Long, and Y. Kang, "Topology-aware graph reinforcement learning for dynamic routing in cloud networks," arXiv preprint arXiv:2509.04973, 2025.
7. P. Li, Y. Wang, E. H. Chi, et al., "Hierarchical reinforcement learning for modeling user novelty-seeking intent in recommender systems," arXiv preprint arXiv:2306.01476, 2023.
8. A. Khamaj and A. M. Ali, "Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time," Alexandria Engineering Journal, vol. 95, pp. 164–173, 2024.
9. A. Chen, C. Du, J. Chen, et al., "Deeper insight into your user: Directed persona refinement for dynamic persona modeling," arXiv preprint arXiv:2502.11078, 2025.
10. S. Pan and D. Wu, "Hierarchical text classification with LLMs via BERT-based semantic modeling and consistency regularization," unpublished.
11. C. Liu, Q. Wang, L. Song, and X. Hu, "Causal-aware time series regression for IoT energy consumption using structured attention and LSTM networks," unpublished.
12. Y. Li, S. Han, S. Wang, M. Wang, and R. Meng, "Collaborative evolution of intelligent agents in large-scale microservice systems," arXiv preprint arXiv:2508.20508, 2025.
13. R. Zhang, "AI-driven multi-agent scheduling and service quality optimization in microservice systems," Transactions on Computational and Scientific Methods, vol. 5, no. 8, 2025.
14. W. Xu, J. Zheng, J. Lin, M. Han, and J. Du, "Unified representation learning for multi-intent diversity and behavioral uncertainty in recommender systems," arXiv preprint arXiv:2509.04694, 2025.
15. V. Shakya, J. Choudhary, and D. P. Singh, "IRADA: Integrated reinforcement learning and deep learning algorithm for attack detection in wireless sensor networks," Multimedia Tools and Applications, vol. 83, no. 28, pp. 71559–71578, 2024.
16. W. C. Kang and J. McAuley, "Self-attentive sequential recommendation," Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), pp. 197–206, 2018.
17. D. Jannach and M. Ludewig, "When recurrent neural networks meet the neighborhood for session-based recommendation," Proceedings of the 2017 ACM Conference on Recommender Systems, pp. 306–310, 2017.
18. N. Vij, A. Yacoub, and Z. Kobti, "XLNet4Rec: Recommendations based on users' long-term and short-term interests using Transformer," Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA), pp. 647–652, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.