Transactions on Computational and Scientific Methods | Vo. 4, No. 6, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Semantic User-Advertisement Alignment via Large Language Models for Personalized Recommendation

Ziruo Ke[1*]
[1]University of Washington, Seattle, USA
*Corresponding author: Ziruo Ke, kziruo@gmail.com

Abstract: This paper addresses the limitations of user interest modeling and semantic matching in personalized advertising recommendation by proposing a framework based on large language models. The method first embeds user behavior sequences and advertisement features, mapping discrete and continuous inputs into a unified semantic space to obtain denser and more expressive representations. A multi-head self-attention mechanism is then introduced to capture deep dependencies between users and advertisements, enabling the identification of key interest signals in long behavior sequences. A joint representation layer aligns user and advertisement representations, and cosine similarity is used to measure their matching degree, ensuring that recommendations better reflect real user needs. The entire model is trained end-to-end with a cross-entropy optimization objective to improve ranking quality and click prediction accuracy. Experiments conducted on a public advertising dataset compare the proposed approach with several mainstream methods, using NDCG, MRR, Recall@10, and Precision@10 as evaluation metrics. Results show that the proposed method achieves the best performance on all metrics. Sensitivity experiments further verify the impact of hyperparameter configurations, especially in attention head number, hidden dimension, and sequence window length, demonstrating stability and adaptability under different conditions. These findings confirm that the proposed framework improves performance across multiple metrics and exhibits strong robustness and practical value in complex and dynamic advertising recommendation scenarios.

Keywords: Personalized recommendations, large language models, attention mechanisms, and ad click-through rates

1. Introduction

In the current digital era, the way information is disseminated and the scenarios of content consumption have undergone profound changes. With the improvement of internet infrastructure and the widespread use of mobile devices, users engage daily with massive amounts of information. Traditional advertising strategies can no longer meet such frequent and diverse demands. Advertising recommendation systems have emerged as an effective solution, improving the efficiency of matching advertisements with users through data-driven approaches[1]. However, rule-based or shallow models rely on limited feature representations and often fail to capture the deep semantic relations behind user behavior. As a result, these systems face issues such as homogeneity of results, cold-start challenges, and preference drift. Against this background, how to enhance the intelligence and personalization of advertising recommendations has become a shared concern for both academia and industry[2].

In recent years, the rise of large language models has provided new opportunities for the advertising recommendation domain. These models, trained on massive text corpora, possess strong natural language understanding and generation capabilities. They can model user behavior, interests, and contexts at a deeper level. Compared with traditional feature extraction methods, large language models not only capture explicit preferences but also uncover hidden interests from interaction histories, user reviews, and contextual

information. This cross-modal and context-aware modeling ability makes them powerful in constructing user profiles and aligning advertisements. In dynamic and rapidly changing consumption scenarios, large language models can continuously update and reason, thus offering more flexible and fine-grained personalized recommendations[3].

Personalization has become the core objective in advertising recommendations. Users show highly diverse interest distributions, and the same advertisement may have very different levels of acceptance among individuals. Hence, recommendation systems must adapt to individual characteristics rather than adopt a "one-size-fits-all" approach. Large language model-based recommendation methods address this demand by understanding language expressions, contextual preferences, and behavioral trajectories. They align the semantic spaces of users and advertisements, producing recommendations that better match user needs. This improves both click-through rates and conversion rates. It also increases the efficiency of ad delivery and enhances user experience by reducing information overload and cognitive burden.

Moreover, advertising recommendation is not only a commercial practice but also related to information fairness, user rights, and social value. In an environment of information overload, users seek more targeted and valuable content to maximize their limited attention. Personalized recommendations powered by large language models can reduce redundant information exposure and make information distribution more precise and efficient. This improvement benefits advertisers as well, helping them achieve more effective market reach and brand promotion. In this way, commercial value and user satisfaction form a positive cycle. This dual effect underscores the practical significance of this research area[4].

In conclusion, research on personalized advertising recommendation driven by large language models represents both an inevitable trend in the integration of artificial intelligence with advertising technologies and a crucial path to addressing the challenges of the information age. On the one hand, it advances recommendation technologies toward higher accuracy and stronger robustness through deep semantic understanding and personalized modeling. On the other hand, it enhances advertising value while improving user experience, thereby injecting new momentum into the growth of the digital economy. Such research carries important academic value as well as broad application prospects and social significance.

2. Related work

In the development of advertising recommendation, early studies mainly relied on collaborative filtering and content matching methods. Collaborative filtering analyzed similarities among users and used historical behaviors for recommendations. It had the advantages of simplicity and scalability, but its performance was limited under cold-start and data sparsity conditions[5]. Content matching methods analyzed the correspondence between advertisements and user interest tags, which improved accuracy to some extent. However, these methods depended too heavily on manually constructed features. They failed to capture complex user behavior patterns and contextual associations and could not adapt to the fast-changing advertising environment. Therefore, although these methods laid the foundation for advertising recommendation systems, they still showed significant limitations in meeting personalized and dynamic recommendation demands[6].

With the rise of deep learning, advertising recommendations gradually introduced neural network structures. This shift enabled a transition from shallow feature representation to deep semantic modeling. Typical neural recommendation models embed users and advertisements into a unified vector space and employ multilayer networks to model nonlinear interactions. Such methods enhanced the expressive power of recommendation systems and captured complex behavioral patterns. Sequence modeling approaches were also widely applied, which effectively handled the temporal dependencies in user behaviors. However, traditional deep learning models still had limitations in long sequence modeling and multimodal information fusion. It was difficult to balance accuracy and efficiency. In advertising scenarios, user behaviors are often highly sparse and dynamic, which places higher demands on model robustness[7].

On this basis, attention-based recommendation methods began to receive increasing attention. These methods dynamically assigned weights to highlight key behaviors or features. This allowed the system to identify the most valuable parts of massive interaction data. Self-attention structures were particularly suitable for long sequence modeling. They captured long-range dependencies without fixed windows, improving both flexibility and accuracy. Further research combined sequence modeling with graph-based modeling. By constructing user-advertisement relationship graphs, graph neural networks enabled preference propagation and representation enhancement across nodes. These approaches demonstrated strong modeling capacity in complex interaction environments. Yet, they still faced the challenge of better understanding semantic content and contextual factors.

In recent years, the development of large language models has opened new paths for advertising recommendations. Their strong semantic understanding and generation abilities allow recommendation systems to extract deep preference features from natural language. This enables high-dimensional semantic alignment between user profiles and advertisement texts. Related studies explored combining large language models with traditional recommendation frameworks. Examples included prompt learning, knowledge enhancement, and context modeling to improve performance. These methods not only enhanced recommendation accuracy but also showed advantages in cold-start and cross-domain transfer. Although challenges remain, such as high computational cost and the need for deeper personalization, large language model-driven personalized recommendations have gradually become an important trend. It provides a broad space for future research in this field.

Open-source recommendation platforms and graph-aware models provide additional infrastructure for this line of research. Unified recommendation frameworks improve reproducible evaluation and modular algorithm comparison [8], while knowledge graph convolution, selective state-space modeling, and interest-unit organization extend personalized recommendation through relational reasoning, efficient sequence representation, and product-level behavioral structure [9], [10], [11].

Recent studies on intelligent infrastructure further show how learning systems can remain stable under dynamic, sparse, and resource-constrained conditions. Game-theoretic reinforcement learning has been applied to competitive cloud resource allocation [12], and self-supervised representation learning with structural dependency modeling has supported early risk detection in distributed systems [13]. Multi-scale convolution and adaptive attention have also been used for backend latency forecasting [14], while graph neural network-driven anomaly detection, secure parameter aggregation, lightweight edge-cloud modeling, and energy-budget task scheduling emphasize robust feature learning and deployment-aware optimization across cloud and edge-cloud environments [15], [16], [17], [18].

These developments extend the methodological base of personalized advertising recommendation through semantic robustness, structured interaction modeling, and decision reliability. Masked language modeling strengthens semantic matching and text robustness [19], differentiable mixed-integer optimization provides a way to connect social graph modeling with node-level decision-making [20], structure-aware multi-scale behavioral learning supports anomaly identification in complex systems [21], and causal inference with counterfactual risk estimation improves the interpretability of automated judgment under uncertainty [22]. Together, these studies broaden the technical foundation for building recommendation models that can align user interests, advertisement semantics, and deployment constraints.

3. Method

The overall framework, centered around a large language model and incorporating structured modeling techniques from recommendation systems, builds an end-to-end personalized ad recommendation algorithm. First, for a user's behavior sequence, let their interaction records be a vector set U and the ad content be a vector set A. Both are mapped into a unified semantic space through an embedding layer. The model architecture is shown in Figure 1.

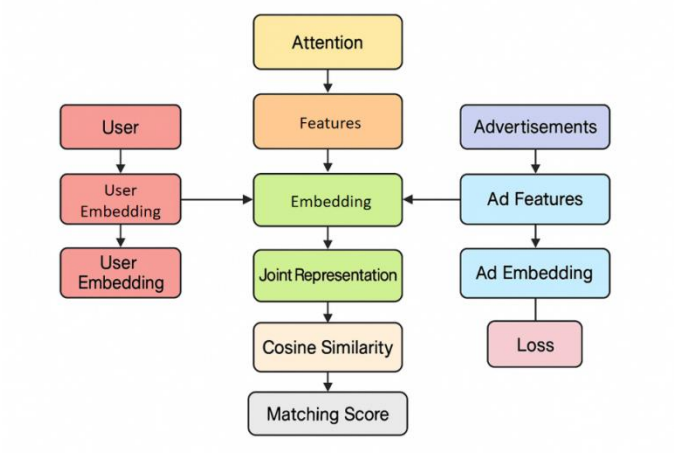

Figure 1. Overall model architecture

The embedding process can be expressed as:

$$h_u = E_u(x_u), \quad h_a = E_a(x_a)$$

x_u represents the user feature input, x_a represents the ad feature input, E_u and E_a are the embedding functions of the user and ad, respectively, and the outputs h_u and h_a are both low-dimensional dense representations. Through this process, the model can provide a foundation for subsequent semantic alignment. In the semantic modeling stage, an attention-based structure is introduced to capture the deep connection between user behavior and advertising content. The attention weight is calculated as follows:

$$\alpha_{ij} = \frac{\exp((h_u^i W_Q)(h_a^j W_K)^T)}{\sum_k \exp((h_u^i W_Q)(h_a^j W_K)^T)}$$

Where W_Q, W_K is a trainable projection matrix, and α_{ij} represents the attention weight between the i-th user behavior and the j-th ad feature. Through this mechanism, the model can adaptively highlight the feature dimensions that contribute most to the prediction, thereby enhancing the discriminability of the representation. The attention-weighted results are then input into the feedforward network to form a joint representation:

$$z_{ua} = \sum_i \alpha_{ij}(h_a^j W_V)$$

In the semantic alignment stage, it is necessary to measure the degree of match between the user representation and the ad representation. For this purpose, a similarity function is introduced, and the commonly used definition is cosine similarity:

$$S(u,a) = \frac{h_u \cdot h_a}{\| h_u \| \| h_a \|}$$

$S(u,a)$ represents the semantic similarity between the user and the ad. The larger the value, the stronger the preference consistency between the two, and the higher the relevance of the recommendation.

Finally, to achieve end-to-end optimization, an objective function based on cross-entropy is adopted:

$$L = -\sum_{(u,a)\in D} [y_{ua} \log \sigma(S(u,a)) + (1 - y_{ua}) \log(1 - \sigma(S(u,a)))]$$

Where D represents the training sample set, y_{ua} represents the actual label indicating whether the user clicked the ad, and $\sigma(\cdot)$ represents the Sigmoid function. By minimizing the loss function, the model continuously adjusts its parameters to make recommendations more aligned with the user's underlying preferences. This approach fully integrates the semantic understanding capabilities of large language models with the personalized modeling capabilities of recommendation systems, achieving a better user experience and greater commercial value in ad recommendation tasks.

The embedding, attention, and similarity modules are therefore organized as a coupled representation pipeline. Structural dependency modeling [13] supports the construction of behavior-advertisement embeddings, multi-scale attention and forecasting designs [14] motivate adaptive weighting across interaction windows, masked language modeling strengthens text robustness in advertisement descriptions [19], graph decision modeling informs relation-aware preference propagation [20], and causal risk estimation provides a useful basis for calibrating ranking scores under uncertain user feedback [22].

4. Experimental Results

4.1 Dataset

This study adopts the Criteo Display Advertising Challenge dataset as a unified data source. The dataset contains hundreds of millions of advertising impression logs. It covers contextual features such as display time, position, device, and region. It also provides anonymized discrete and continuous features from both the user and advertisement sides, together with a binary label indicating whether a click occurred. The dataset is large in scale, rich in feature types, and stable in distribution over time. It reflects key characteristics of real advertising scenarios, including sparsity, long-tail distribution, and strong contextual dependence. It is therefore widely used as a standard benchmark for personalized advertising recommendation and click-through rate modeling.

The raw data consists of 13 numerical features and 26 high-cardinality categorical features. The categorical features have been anonymized by hashing to protect privacy and business confidentiality. Common practices include filling missing values and applying bucket-based normalization for numerical features. For categorical features, frequency-based truncation and transformations such as hashing or target encoding are often applied to mitigate dimensional explosion and sparsity caused by high cardinality. The dataset is typically split into training, validation, and test sets based on chronological order, which allows evaluation of generalization performance under distribution shift. The design of fields and the granularity of records support multiple modeling perspectives, ranging from point-level click modeling to session-level sequence modeling.

In terms of compliance, the dataset has been anonymized and contains no information that can directly identify individuals. It is suitable for algorithmic research under privacy protection requirements. To further reduce potential risks, studies usually follow common principles such as minimizing identifiability and limiting usage. Sensitive contextual features such as geography and device type are often discretized or perturbed. Derived features are used and published with care to avoid any possibility of reverse identification. These measures ensure a balance between practical usability and privacy protection and provide a reliable and compliant foundation for subsequent personalized modeling.

4.2 Experimental Results

This paper also gives the comparative experimental results, as shown in Table 1.

Table 1. Comparative experimental results

Model	NDCG	MRR	Recall@10	Precision@10
Recbole[8]	0.415	0.392	0.528	0.267
RAKCR[9]	0.438	0.407	0.551	0.281

Model	NDCG	MRR	Recall@10	Precision@10
Mamba4rec[10]	0.462	0.423	0.576	0.298
IU4Rec[11]	0.479	0.439	0.593	0.312
Ours	0.516	0.471	0.631	0.341

From the overall results, the proposed method outperforms existing baseline models on all four core metrics. This demonstrates that large language model-driven personalized advertising recommendation has clear advantages in semantic understanding and user behavior modeling. Both NDCG and MRR show improvements in the overall ranking quality and the relevance of top-ranked results. This indicates that the model captures users' potential interests more effectively, which enhances the alignment between advertisements and user preferences.

A closer look at Recall@10 shows that the improvement in recall is the most significant. Compared with traditional deep models, the proposed framework achieves an increase of 0.038. This suggests that the method has stronger advantages in covering diverse user interests. It can identify and recommend more candidate advertisements that meet real user needs. As a result, it reduces omissions and improves the comprehensiveness of recommendations. This characteristic is especially important in advertising scenarios, where user interests are often highly sparse and dynamic.

On the Precision@10 metric, the proposed method also achieves the best performance. This shows that the top 10 recommended advertisements have higher relevance and less noise. Compared with traditional methods, the proposed approach maintains the improvement in recall while also achieving higher precision. It avoids the common trade-off in recommendation systems where broader coverage leads to lower accuracy. This result demonstrates the effectiveness of large language models, combined with semantic modeling and attention mechanisms, in extracting user interests and matching advertisements.

In summary, the experimental results indicate that the proposed framework not only surpasses existing methods in overall ranking and top-ranked recommendation quality but also achieves a better balance between coverage and precision. This means that the introduction of large language models into advertising recommendations can provide users with more personalized and relevant content, while also delivering higher conversion value for advertisers. These findings offer strong empirical support for the intelligent development of future advertising recommendations.

This paper also presents an experiment on the sensitivity of the number of attention heads to NDCG@10, and the experimental results are shown in Figure 2.

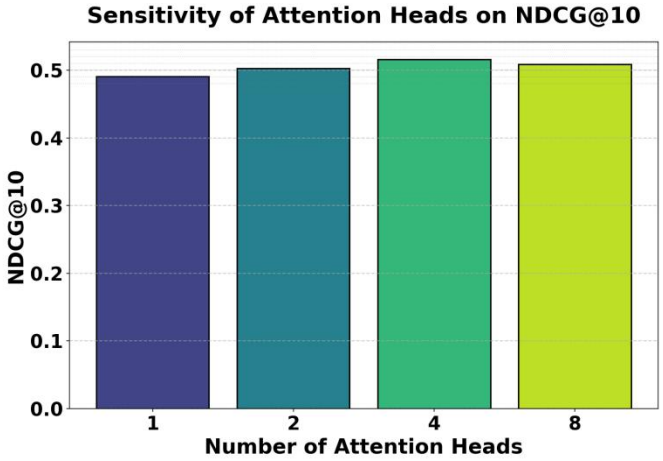

Figure 2. Experiment on the sensitivity of the number of attention heads to NDCG@10

The results show that as the number of attention heads increases, the model achieves steady improvement in NDCG@10. When the number of heads increases from 1 to 4, the improvement is more obvious. This indicates that the multi-head mechanism captures diverse relationships between user behaviors and advertisement features more effectively, giving the model an advantage in ranking performance. Multi-head attention provides parallel modeling ability across different representation spaces, which allows better modeling of complex user interest patterns.

When the number of attention heads further increases from 4 to 8, the gain in NDCG@10 becomes smaller and even shows a slight decline. This suggests that too many heads may introduce redundant information. The interference makes it harder for the model to focus on key interest signals, limiting the improvement of overall ranking performance. This result highlights the need to balance the number of attention heads and model complexity, ensuring accuracy while avoiding unnecessary resource consumption.

By comparing performance with different numbers of attention heads, it is clear that a moderate number improves the ability of the model to capture user interests. Too few heads limit the capacity to model diverse features, making it difficult to fully align user and advertisement semantics. Too many heads create redundancy, which weakens the model's focus and reduces the relevance of recommendations.

In conclusion, these results confirm the important role of attention mechanisms in personalized advertising recommendations. They also emphasize the critical influence of proper hyperparameter selection on final performance. Sensitivity analysis of attention head numbers guides model deployment in real applications. It ensures recommendation quality while considering efficiency and resource use, and lays a solid foundation for optimizing large language model-driven recommendation frameworks.

This paper also gives the influence of the sequence window on experimental results, emphasizing how the length of the user interaction sequence considered by the model can affect its ability to capture temporal dynamics and contextual dependencies in recommendation tasks. The sequence window defines the scope of historical behaviors that are taken into account when generating predictions, thereby determining how much past information contributes to the construction of user preference profiles. A shorter window may focus on more recent interactions and highlight short-term interests, while a longer window incorporates a broader range of behaviors, potentially capturing long-term patterns but also introducing noise or redundant signals. By varying the sequence window, this experiment is designed to analyze the trade-offs between precision, recall, and computational cost, while also providing a clearer understanding of how temporal context influences the learning process.

The motivation for investigating the sequence window lies in the dynamic nature of advertising recommendation environments, where user interests can evolve rapidly over time but may also contain stable, long-term preferences. Determining an appropriate sequence window size is therefore critical to achieving a balance between responsiveness to immediate changes and stability in long-term modeling. Moreover, the choice of window length has direct implications for model complexity and runtime performance, as larger windows increase both data processing demands and training overhead. Through this sensitivity analysis, the study aims to highlight the importance of tailoring sequence modeling strategies to the unique requirements of advertising recommendation systems. The experimental results are shown in Figure 3.

From the overall results, as the sequence window length increases, all metrics show a certain degree of improvement. The most significant gain occurs when the window length increases from 5 to 20. This indicates that an appropriate window length can capture more contextual information in user behavior sequences, thereby improving the overall quality of recommendation ranking. The improvement in NDCG@10 and MRR demonstrates the enhanced ability of the model to identify relevant advertisements and optimize top-ranked results.

For the Recall@10 metric, the recall rate gradually increases as the window length grows. This shows that longer sequences provide the model with richer contextual cues, enabling it to cover more potentially

relevant advertisements. This is important for advertising recommendation tasks, as user interests are often distributed across multiple dimensions. A longer history reduces omission and enhances the comprehensiveness of recommendations.

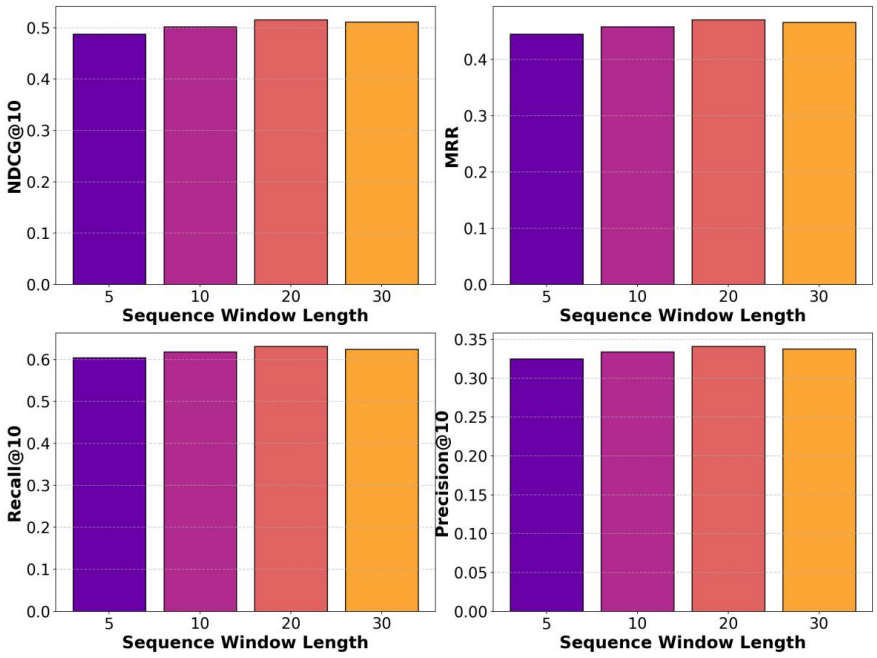

Figure 3. The impact of sequence windows on experimental results

The results for Precision@10 reveal that when the window length becomes too long, the improvement in precision slows down and even slightly decreases at a length of 30. This suggests that overly long sequences may introduce redundancy or noise, weakening the model's focus on key interest signals. In recommendation scenarios, excessive irrelevant interactions may reduce the probability of highly relevant advertisements being ranked at the top. Therefore, the window length needs to be reasonably constrained to maintain recommendation relevance.

In summary, these experimental results confirm the significant impact of sequence window length on model performance. They show that a moderate window length can achieve a balance between coverage and precision. Short sequences fail to capture sufficient context, while long sequences may introduce noise. Only with a proper window length can large language models maximize their effectiveness in advertising recommendation tasks and provide more robust This paper also gives the impact of the hidden dimension size on the experimental results, focusing on how variations in the representational capacity of the model influence its ability to capture complex semantic patterns in user behaviors and advertisement features. The hidden dimension is a crucial hyperparameter in deep learning–based recommendation systems, as it determines the granularity and richness of the learned embeddings. Smaller hidden dimensions may limit the expressive power of the model, potentially constraining its ability to represent nuanced relationships between users and ads, while larger hidden dimensions offer greater capacity to encode fine-grained interactions but may also introduce risks such as overfitting or unnecessary computational overhead. By systematically adjusting the hidden dimension size, the experiment is designed to explore this trade-off and to evaluate how different settings impact the balance between efficiency and effectiveness in recommendation tasks.

The inclusion of this analysis is motivated by the need to understand the sensitivity of recommendation frameworks to architectural choices that directly affect both performance and scalability. In large-scale advertising systems, the cost of training and serving models grows with the dimensionality of the embeddings, making it essential to identify an optimal configuration that maximizes recommendation

quality while maintaining computational feasibility. Examining the hidden dimension size also provides valuable insights into how the proposed framework adapts to different levels of model complexity and resource availability, ensuring that it can be effectively deployed in diverse environments with varying constraints. The experimental results are shown in Figure 4.

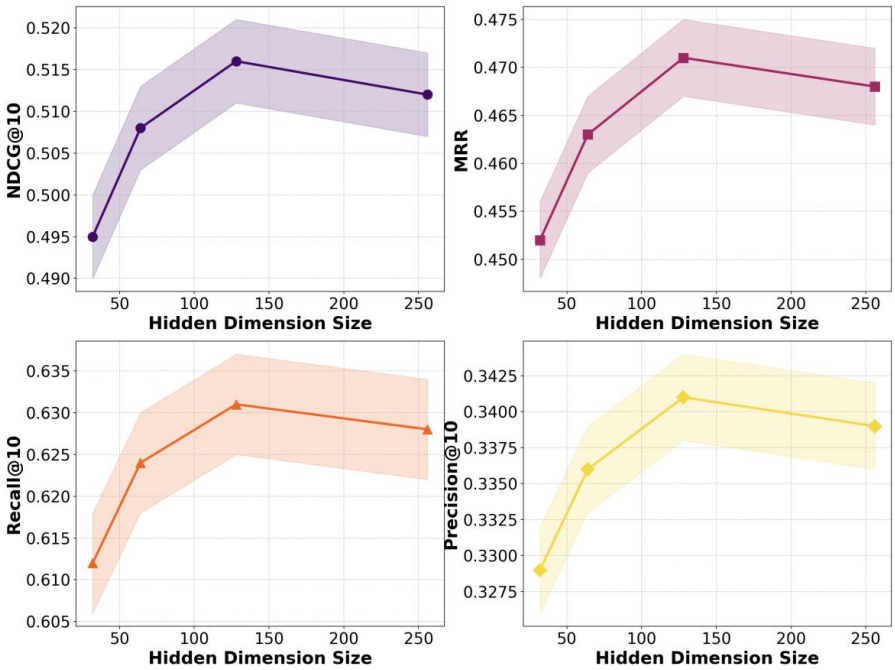

Figure 4. The impact of hidden dimension size on experimental results

From the results, it can be observed that as the hidden dimension increases from 32 to 128, the model achieves significant improvements across all metrics. The gains in NDCG@10 and MRR are particularly notable. This indicates that higher-dimensional representations provide a richer semantic space, enabling the model to capture more fine-grained user-advertisement preference relations. Such improvement directly enhances the ranking quality and the relevance of top recommendations.

In Recall@10, the performance continues to rise as the dimension grows to 128. This shows that stronger representation ability helps the model cover more potentially relevant advertisements. Higher-dimensional embeddings reduce omission and improve the comprehensiveness of recommendations. However, when the dimension increases further to 256, the improvement slows and even shows a slight decline. This reflects that very high dimensions may introduce redundant features, making it harder for the model to maintain focus on key interest signals.

The trend of Precision@10 is similar to that of Recall@10. Precision reaches its best at a moderate dimension, but slightly drops at very high dimensions. This suggests that an excessively large hidden space may increase noise and the risk of overfitting, leading to reduced relevance of recommendations. Therefore, larger hidden dimensions are not always better. A balance between expressive capacity and robustness is required.

In summary, these results highlight the critical role of hidden dimension size in personalized advertising recommendation tasks. A suitable dimension can significantly enhance semantic modeling and recommendation performance, while too small or too large dimensions weaken the model. Sensitivity experiments provide valuable guidance for model design, allowing large language model–based recommendation frameworks to achieve a better trade-off between efficiency and accuracy, and offering stronger support for large-scale advertising applications.

This paper also presents a sensitivity experiment on Recall@10 in scenarios with increasing cold start ratios, aiming to investigate how the performance of the proposed recommendation framework adapts when the

amount of available user interaction history becomes limited. The cold start problem is a well-recognized challenge in recommendation systems, where insufficient user data leads to difficulties in accurately capturing individual preferences and providing high-quality recommendations. By gradually adjusting the proportion of users without sufficient historical behavior, this experiment is designed to simulate realistic application environments in which new users continuously enter the system and interact with advertisements. Such a setup allows for a more comprehensive understanding of how robust the framework is under different levels of data sparsity and user initialization conditions.

The motivation for incorporating this experiment lies in the need to evaluate the generalization ability and stability of recommendation algorithms in dynamic and evolving ecosystems. In large-scale advertising platforms, user profiles are rarely static, and the presence of cold start users is inevitable as the platform grows and diversifies. Thus, testing Recall@10 under varying cold start ratios provides critical insights into the algorithm's capacity to maintain recommendation coverage and adapt to scenarios characterized by incomplete or sparse user behavior sequences. This sensitivity analysis not only deepens the understanding of algorithmic behavior under challenging conditions but also highlights the importance of designing recommendation frameworks that remain effective across heterogeneous user groups.

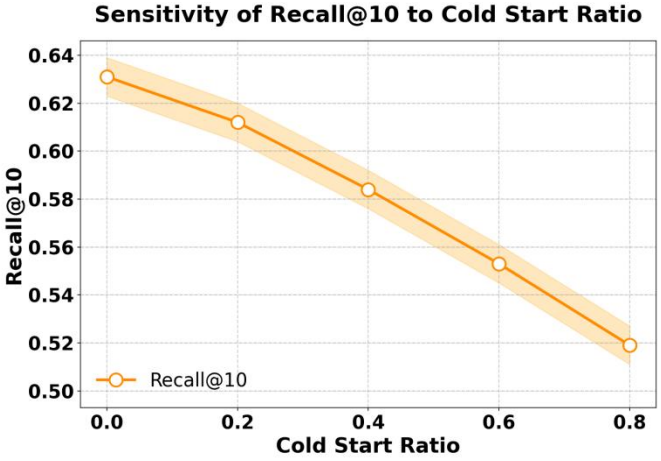

Figure 5. Sensitivity experiment on Recall@10 in scenarios with increasing cold start ratio

From the experimental results, it can be seen that Recall@10 decreases significantly as the cold-start ratio increases. This indicates that in advertising recommendation tasks, when the system lacks sufficient user history, the model's ability to capture user interests is limited. As a result, the number of relevant advertisements that can be recalled is reduced. The sparsity of data caused by cold-start directly affects semantic modeling and recommendation performance.

When the cold-start ratio is low, Recall@10 remains at a high level. This shows that with richer interaction records, the model can accurately capture user interests. At this stage, the recommendations achieve strong coverage and relevance. It also demonstrates that the semantic understanding ability of large language models can be fully utilized when sufficient data are available, providing reliable support for recommendation tasks.

As the cold-start ratio continues to increase, the decline of Recall@10 accelerates. The drop becomes more significant after 0.6. This suggests that when most users lack historical data, the model struggles to rely only on semantic matching and contextual modeling to fill the information gap. The bottleneck in covering potentially relevant advertisements highlights the need for additional mechanisms, such as cross-domain knowledge transfer or content-based modeling.

Overall, these results highlight the sensitivity of recommendation performance to the cold-start problem. They emphasize that relying only on historical interactions may be insufficient to support recommendation quality. In large language model–based frameworks, how to capture latent user interests and improve recall

in cold-start conditions remains an important challenge. This finding provides practical guidance for optimizing large-scale personalized recommendation systems in real-world applications.

5. Conclusion

This study focuses on personalized advertising recommendations driven by large language models. It systematically explores innovative methods in semantic modeling, user interest representation, and advertisement content matching. By combining the deep semantic understanding ability of large language models with the personalized modeling needs of recommendation systems, the study demonstrates that such integration can effectively address challenges such as homogeneity, cold-start, and data sparsity. As a result, the overall quality of advertising recommendations and the user experience are improved. This conclusion not only confirms the applicability of large language models in recommendation systems but also provides new theoretical support for the development of recommendation technology.

In the experimental results, the proposed framework achieves superior performance in ranking quality, coverage, and precision. The key improvement lies in the ability of large language models to align semantic spaces and to integrate user behavior, advertisement features, and contextual information. This enables more fine-grained personalized recommendations. Compared with traditional collaborative filtering or shallow deep learning methods, the framework captures more complex interest dynamics and cross-modal relations, making the recommendations more accurate. This provides strong evidence for the intelligent delivery of advertisements and highlights the unique value of large language models in complex recommendation tasks.

From an application perspective, the significance of this research is not limited to advertising recommendations. In real-world scenarios, users face increasingly complex information environments, and recommendation systems must provide highly relevant content within a limited time and space. The proposed framework reduces redundancy and ineffective recommendations, thereby enhancing user satisfaction and interaction experience. For advertisers, this approach improves the precision and conversion rates of advertisement delivery, leading to greater commercial returns. Thus, the study demonstrates practical value in both user experience optimization and business outcomes.

More importantly, the proposed approach has cross-domain potential. The logic of applying large language models in personalized recommendations can also be transferred to education, healthcare, and finance. In these areas, understanding user needs and providing accurate services are also core challenges. Therefore, this study not only advances advertising recommendation but also lays a methodological foundation for building broader intelligent recommendation systems. Such cross-scenario scalability underscores the potential and importance of large language model – driven recommendation frameworks in the future artificial intelligence ecosystem.

References

[1] Wu L, Zheng Z, Qiu Z, et al. A survey on large language models for recommendation[J]. World Wide Web, 2024, 27(5): 60.

[2] Li F, Li Y, Liu Y, et al. LEADRE: Multi-Faceted Knowledge Enhanced LLM Empowered Display Advertisement Recommender System[J]. arXiv preprint arXiv:2411.13789, 2024.

[3] Lin J, Dai X, Xi Y, et al. How Can Recommender Systems Benefit from Large Language Models: A Survey[J]. ACM Transactions on Information Systems, 2024.

[4] Awad A, Roberts D, Dolev E, et al. adSformers: Personalization from short-term sequences and diversity of representations in Etsy ads[J]. arXiv preprint arXiv:2302.01255, 2023.

[5] Li Y, Zhu J, Liu W, et al. Pear: Personalized re-ranking with contextualized transformer for recommendation[C]//Companion Proceedings of the Web Conference 2022. 2022: 62-66.

[6] Yu P, Xu Z, Wang J, et al. The application of large language models in recommendation systems[C]//International Conference on Artificial Intelligence and Machine Learning Research (CAIMLR 2024). SPIE, 2025, 13635: 103-112.
[7] Li L, Zhang Y, Liu D, et al. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024.
[8] Zhao W X, Mu S, Hou Y, et al. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms[C]//proceedings of the 30th acm international conference on information & knowledge management. 2021: 4653-4664.
[9] Cui Y, Yu H, Guo X, et al. RAKCR: Reviews sentiment-aware based knowledge graph convolutional networks for Personalized Recommendation[J]. Expert Systems With Applications, 2024, 248: 123403.
[10] Liu C, Lin J, Wang J, et al. Mamba4rec: Towards efficient sequential recommendation with selective state space models[J]. arXiv preprint arXiv:2403.03900, 2024.
[11] Sanner S, Balog K, Radlinski F, et al. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences[C]//Proceedings of the 17th ACM Conference on Recommender Systems. 2023: 890-896.
[12] Hsieh C Y. Game-Theoretic Reinforcement Learning for Stable Competitive Resource Allocation in Dynamic Cloud Environments[J], 2024, 4(1).
[13] Ying T, Ding C. Self-Supervised Unified Representation Learning with Structural Dependency Modeling for Early Risk Detection in Distributed Systems[J]. , 2024, 4(2).
[14] Pi S. Intelligent Backend Latency Forecasting with Multi-Scale Convolution and Adaptive Attention Mechanisms[J], 2024, 3(5).
[15] Shawulieti S. Enhancing Cloud Security with Graph Neural Network-Driven Anomaly Detection in Multi-Tenant Environments[J], 2024, 4(2).
[16] Zhang S. Secure Parameter Aggregation and Distributed Anomaly Detection in Cross-Cloud Data Center Networks[J], 2024, 3(4).
[17] Xu S. Efficient Anomaly Detection in Distributed Edge-Cloud Systems Using Lightweight Modeling[J], 2024, 4(3).
[18] Sun J. Reinforcement Learning-Based Task Scheduling Under Energy Budget Constraints in Edge-Cloud Systems[J], 2024, 4(4).
[19] Bian Z, Su X. Enhancing Semantic Matching and Text Robustness through Masked Language Modeling[J], 2024, 4(1).
[20] Ren X. A Unified Framework for Social Graph Modeling and Node Decision-Making via Differentiable Mixed-Integer Optimization[J], 2024, 3(5).
[21] Zhou Y. Structure-Aware Multi-Scale Behavioral Learning for Distributed System Observability and Anomaly Identification[J], 2024, 4(3).
[22] Lin S. Automatic Audit Judgment Using Causal Inference and Counterfactual Risk Estimation[J], 2024, 4(3).