

User Feedback Alignment for LLM-powered Exploration in Large-scale Recommendation Systems

Jianling Wang¹, Yifan Liu², Yinghao Sun³, Xuejian Ma², Yueqi Wang², He Ma², Steven Su²,
Ed H. Chi¹, Minmin Chen¹, Lichan Hong¹, Ningren Han², Haokai Lu¹

¹Google DeepMind ²YouTube ³Google Labs

{jianlingw, yifanliu, sunmo, xuejianma, yueqi, htm, susteven,
edchi, minminc, lichan, peterhan, haokai}@google.com

Abstract

Exploration, the act of broadening user experiences beyond their established preferences, is challenging in large-scale recommendation systems due to feedback loops and limited signals on user exploration patterns. Large Language Models (LLMs) offer potential by leveraging their world knowledge to recommend novel content outside these loops. A key challenge is aligning LLMs with user preferences while preserving their knowledge and reasoning. While using LLMs to plan for the next novel user interest, this paper introduces a novel approach combining hierarchical planning with LLM inference-time scaling to improve recommendation relevancy without compromising novelty. We decouple novelty and user-alignment, training separate LLMs for each objective. We then scale up the novelty-focused LLM’s inference and select the best-of-n predictions using the user-aligned LLM. Live experiments demonstrate efficacy, showing significant gains in both user satisfaction (measured by watch activity and active user counts) and exploration diversity.

1 Introduction

Large Language Models (LLMs) present a significant opportunity to revolutionize recommendation systems (Wu et al., 2024), due to their powerful reasoning, planning, and world knowledge capabilities. Traditional recommendation backbones, such as collaborative filtering and content-based methods, typically suggest items by identifying similar users based on past interactions, which often reinforce existing preferences and perpetuate feedback loops (Chaney et al., 2018; Mansoury et al., 2020). LLMs can overcome these limitations by leveraging their vast world knowledge to generate recommendations that go beyond a user’s historical interactions, introducing novel and diverse items, to drive long-term user engagement (Chen, 2021) and reduce reliance on past behavior alone.

Among recent advancements leveraging LLMs for recommendation systems (Bao et al., 2023; Lin et al., 2024a; Wang et al., 2024a; Christakopoulou et al., 2023), the hierarchical planning paradigm (Wang et al., 2024b) stands out a promising and *deployable* approach that combines LLMs for high-level guidance with traditional recommenders for efficient item-level serving. As this solution has been adopted in industry, the next challenge is integrating real-world human feedback into the LLM. While human feedback is key to optimizing LLMs (Ouyang et al., 2022), systematically incorporating it into recommendation systems remains an under-explored area, offering both challenges and opportunities for future research.

Effectively using real-world human feedback is challenging because recommendation systems rely on noisy implicit signals (e.g., clicks or dwell time) instead of explicit comparative judgments (e.g., side-by-side comparisons). This make it hard to translate such feedback into robust training objectives for LLMs that align with users’ true preferences. More importantly, balancing novelty and relevance, two usually competing objectives, is crucial for exploration in recommendation systems as relevant novel contents drive sustained user satisfaction. Initial experiments with the hierarchical planning (Wang et al., 2024b) framework, using an LLM as a novelty model to identify novel interest cluster and subsequently retrieve relevant items, demonstrated the potential of this approach. However, aligning the novelty model’s predictions with user preferences remains challenging. Directly fine-tuning with more users’ interaction history data yielded neutral results, also raised concerns about memorization and loss of novelty. Attempts at RLHF (Ouyang et al., 2022) with a reward model also proved unsuccessful as it undermine the controlled generation capability (see in Sec. 3).

To address these challenges, we propose a novel, decomposed approach that leverages two special-

ized LLM models for high-level planning: a novelty model and an alignment model. To balance novelty and relevance, the alignment LLM is trained specifically to rate the novelty model’s predictions based on observed user feedback. This separation allows for the independent optimization of novelty generation and preference alignment. Moreover, to further improve the system’s ability to generate relevant novel predictions, we scale inference-time compute by generating multiple independent predictions from the novelty model using a high temperature setting. The alignment model then acts as a selector, choosing the most user-aligned outputs from the novelty model. This combination of specialized models, collective user behaviors as training signal, and repeated sampling significantly increases the likelihood of generating recommendations that are both novel and relevant.

In summary, this paper presents the following key contributions: (1) **Collective User Feedback Alignment:** We introduce an LLM-based alignment model specifically trained to evaluate the novelty model’s predictions based on collective user behaviors. By aggregating implicit signals like clicks and dwell time across user clusters, we enable the system to learn user preferences with reduced noise and bias. (2) **Inference-Time Scaling:** We demonstrate the effectiveness of repeated sampling at inference time, allowing the alignment model to select the most relevant predictions from a diverse set of candidates generated by the novelty model, thereby improving exploration. (3) **Decomposed Novelty and Preference Modeling:** We propose a novel paradigm that decouples novelty generation and preference modeling into two specialized LLMs. This separation allows for independent optimization of each objective, resulting in a significantly improved operating curve for user interest exploration by directly addressing the core challenge of balancing novelty with relevance through specialized models. The system is **deployed** on a commercial short-form video recommendation platform serving billions of users.

2 Related Work

LLMs for Recommendation Systems. The advances in LLM capabilities recently has drawn a lot of attention to their potential in recommendation systems (Bao et al., 2023; Dai et al., 2023; Geng et al., 2023; Hou et al., 2023; Li et al., 2023; Liu et al., 2023a). One promising direction involves

augmenting traditional recommendation models with LLM-powered feature engineering, including supplementary textual features or embeddings encoded world knowledge (Xi et al., 2024; Ren et al., 2024; Liu et al., 2023b). Another approach focuses on directly generating recommendations using LLMs. For instance, Hou et al. and Gao et al. have experimented with prompting off-the-shelf LLMs to produce ranked lists of recommendations. Meanwhile, there are also work involves fine-tuning LLMs (Singh et al., 2024; Bao et al., 2023; Lin et al., 2024b) to better align them with the recommendation domain, whether through incorporating domain-specific knowledge, generating new tokens, or predicting user preferences for specific user-item pairs. However, few of these methods are truly equipped to handle query-per-second (QPS) requirements of real-time applications. (Wang et al., 2024a) addresses this by employing LLMs as data augmentation tools for conventional recommendation systems during training, thereby boosting performance without incurring additional serving costs.

Recommendation Exploration. Improving user interest exploration is key to broadening preferences and fostering long-term engagement (Chen et al., 2021; Chen, 2021; Su et al., 2024). However, a key challenge lies in the inherent closed-loop nature of existing recommendation systems (Chaney et al., 2018; Mansoury et al., 2020; Chen et al., 2021). Training data is primarily derived from past user-item interactions, limiting the system’s ability to explore truly novel interests. While methods like PIE (Mahajan et al., 2023) offer improvements through user-creator affinity and online bandit formulations, they remain confined by the system’s internal knowledge (Chen et al., 2021). Building on the LLM-powered hierarchical architecture of (Wang et al., 2024b), which guides user interest exploration at the cluster level, we focus on enhancing its performance through user feedback alignment. Our work represents an early investigation of effective user feedback signals and their integration into LLMs for recommendation systems.

3 Preliminaries

Hierarchical Planning Paradigm. In the hybrid hierarchical planning paradigm (Wang et al., 2024b), LLMs focus on high-level planning by predicting novel user interests at the interest cluster level. Interest clusters are topically coherent item

clusters generated from item metadata and content embedding (Chang et al., 2024). To grant LLM with domain knowledge of our system, we finetuned the LLM using the novel interest transition patterns mined from users’ interaction history.

As illustrated in Figure 1, during the high-level planning, given a user’s recent interaction history, represented as a sequence of K clusters S_u , with $S_u \subseteq \mathcal{C}$ and $|S_u| = k$, the LLM predicts the next novel cluster C_n for this user. Because on-line serving the LLM for a billion-user system is prohibitively costly, we pre-compute and store potential next interest transitions for all combinations of sampled k clusters $\mathcal{S} = \{S \mid S \subseteq \{C_1, C_2, \dots, C_N\}, |S| = k\}$. During online serving, a user’s history is mapped to the corresponding pre-computed novel interest through looking up the precomputed interest transitions. At the lower level, a conventional, transformer-based sequential recommender backbone handles the computationally intensive task of item-level selection. However, instead of searching the entire item space, the backbone is constrained to recommend items only within the novel interest cluster C_n identified by the LLM. This constraint combines the personalization capabilities of the backbone with the novelty-seeking behavior of the LLM, leading to a personalized recommendation experience enriched with serendipitous discoveries.

We’ve launched this user interest exploration paradigm to the production recommendation system, which resulted in a rare combination of high novel item ratio and user satisfaction gain. The lightweight finetuning (<8k training examples) played a key role in preserving the LLM’s own knowledge while teaching it some understanding of our users’ interaction patterns.

Limitation. The lightweight finetuning, however, has limitations: 1) The 8k training examples represented a limited view of the behavior of our large user base. 2) For the cluster combinations that are hard to reason, LLM has low prediction confidence, indicated by the novel interest predictions that don’t have a logical connections to the users’ existing interests. This hurts the relevancy of the recommendation thus user satisfaction.

To improve the relevancy of the novel interest prediction, we initially try to increase the number of training examples, but novelty metrics are not sensitive to the change in A/B testing. Furthermore, due to the LLM’s tendency to repeat training data

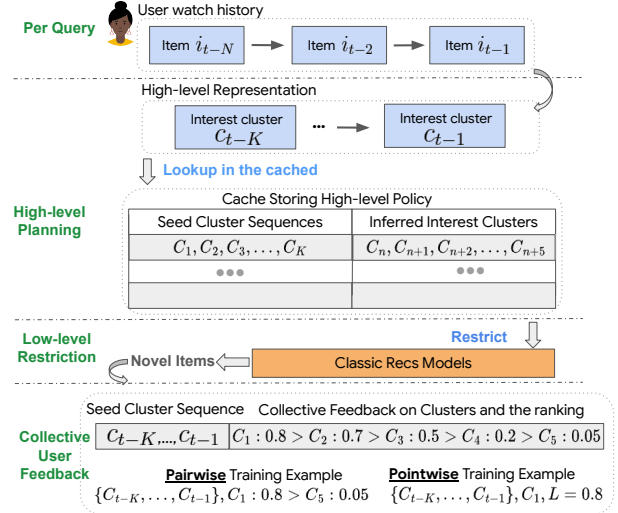


Figure 1: Hierarchical planning paradigm.

(our analysis showed a 40% chance of repetition during inference), scaling to more training example mined from the user history risked reinforcing the system feedback loop, impairing LLM’s ability to make novel recommendations.

To align the LLM with user preference without amplifying the system feedback loop, we leverage live-traffic users’ feedback to LLM’s own recommendations, such as clicks, dwell time and repeated interaction, which is independent of the system behavior. We first tried the classic RLHF setup: RL finetune the novelty LLM directly with a reward model trained with user preference. However, this always resulted in the model quickly collapsing: 1) loss of controlled generation: after 5k steps, the LLM’s chance of predicting in the correct format drops from 99+% to 2%; 2) Reward hacking: the model learned the high reward words, e.g. cat, BTS, toys, etc, and frequently predicts those words. While RLHF is effective in free form text generation in conversation settings, it turns out for structured tasks with strict format and content vocab requirement, a reward model is insufficient and cannot capture the nuanced task requirements and guide the RL finetuning process accordingly.

4 Method

To address the challenges in the classic RLHF, we introduce an inference-time scaling method (Brown et al., 2024) with a decoupled dual-specialization modeling approach. Instead of directly fine-tuning the policy model (i.e., the novelty model responsible for planning the next cluster) through SFT or RLHF, our approach first performs independent

sampling from the novelty model. This generates a diverse set of candidate interest clusters. And subsequently the best-of-n clusters are selected using a separate alignment model trained on collective user feedback based on their likelihood to resonate with users.

This section details our design, demonstrating: (1) the methodology for collecting and transforming implicit user feedback from interactions with the recommendation system into fine-tuning signals for the alignment model; and (2) a top-n selection strategy and inference scaling approach that simultaneously optimizes for both relevance and novelty with minimum latency impact, showcasing its practical applicability in large-scale real-world recommendation systems.

4.1 Preference Alignment on User Feedback

Aggregating Collective Human Feedback. Through per-query logging inside our LLM-powered recommender serving live traffic (detailed in Section 3, ‘Novelty model’ hereafter), we collect users’ preferences on LLM’s predictions. Specifically, for each predicted cluster C_n , we log the cluster sequence $\{C_1, \dots, C_K\}$ used to represent the user, and the user’s feedback on C_n (e.g., positive playback, like, share, skip, etc). We then aggregate the feedback for each $(\{C_1, \dots, C_K\}, C_n)$ pair, resulting in user preference training example denoted as $(\{C_1, \dots, C_K\}, C_n, L_{(1,k),n})$. Here, $L_{(1,\dots,k),n}$ represents the aggregated user feedback score (e.g. like rate, share rate) for this particular interest cluster transition – that is, serving interest cluster C_n to a user with historical viewing pattern represented by $\{C_1, \dots, C_K\}$.

We then post-process the aggregated feedback to: 1) normalize the feedback score, which can be skewed to very small value because the feedbacks, e.g. like, share, etc, are sparse. 2) filter cluster transition pairs with few user feedback. 3) round the feedback score to a fixed interval to account for margin of error in the aggregated stats.

Besides the aforementioned *pointwise training example* $(\{C_1, \dots, C_K\}, C_n, L_{(1,k),n})$, we also tested pairwise training examples: we rank the different C_n for a cluster sequence $\{C_1, \dots, C_K\}$ by the aggregated feedback score, and we create training examples by sampling contrastive C_n pairs as labels. Pairwise training examples require neither normalization nor picking a threshold for positive labels. We can also generate more training (K-

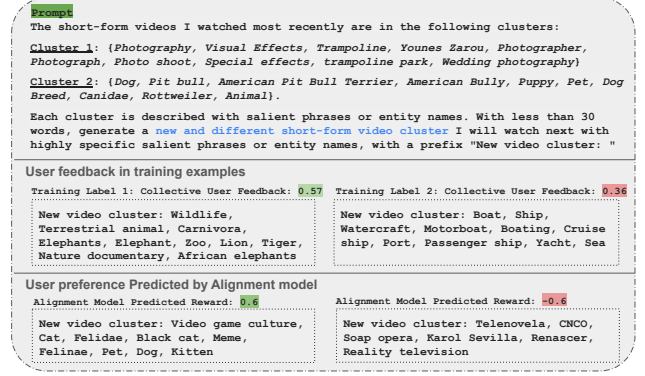


Figure 2: The alignment model trained with collective user feedback can effectively predicts user preference over new labels.

choose-2 vs K) examples per cluster sequence.

Alignment Reward Model Training. To align with collective user feedback, we trained an alignment reward model (‘alignment model’ hereafter) which scores on C_n , using a cross-entropy loss between the prediction and the user’s actual aggregated engagement metric (i.e., positive playback rate). This alignment model is an LLM with the last layer being a linear projection layer.

In Figure 2, we showcase a sample prompt describing users with {photography, Visual Effects, Special effects} and {dogs} interest clusters (assuming $K = 2$). Collectively, those users prefer label 1 {wildlife, nature documentary} over label 2 {boats} as expressed in the feedback scores. Given two new labels, the trained alignment model also effectively assigns high preference score to {cats, video game, internet meme} over less relevant next interest cluster. These intuitive examples demonstrate the feasibility and potential of the alignment training.

4.2 Inference Scaling with Best-of-N User Alignment

We use the user alignment model as a user surrogate to critique the relevancy of the novel clusters predicted by the novelty LLM, which is lightly fine-tuned with users’ interaction histories. To increase the chance of predicting a novel cluster that is more aligned with user preference, we repeatedly and independently sample 5 times more predictions from the novelty LLM with high temperature, and then rank the predictions using the alignment model and pick the top k where k is the number of clusters served by the production system. Because the novelty LLM sampling, reward model scoring, and

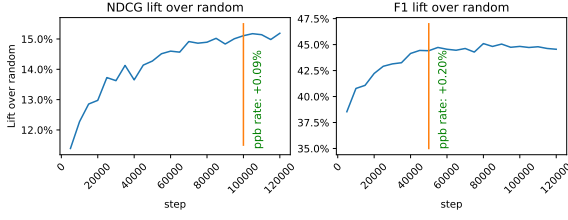


Figure 3: Alignment Model Finetuning and Evaluation.

the best-of-n selection all happens offline, and we serve the same number of clusters in live traffic, there is no latency impact, and the additional cost of scaling up inference is amortized across offline bulk inference runs.

Getting a guarantee of the novelty of the prediction is important for user interest exploration. The repeated sampling of the novelty LLM improves the reasoning quality and maintains the prediction novelty while the alignment model selects the predictions users may prefer. This dual LLM setup avoids the challenge of teaching LLM both novelty and relevancy, which are two competing objectives and may result in catastrophic forget. By reflecting on the novelty prediction using an LLM aligned with user feedback, we improve the exploration efficiency by demoting the predictions that may result in lower user satisfaction.

5 Live Experiments

5.1 Experimental Setup

Our live experiments were conducted on a commercial short-form video recommendation platform serving billions of users. While we employed Gemini (Team et al., 2024) for both the novelty and alignment models, the fine-tuning process and pipeline are designed to be adapted to other LLMs. The high-level planning recommends novel interest clusters based on a user’s historical interest cluster sequence of length $K = 2$, and the system is designed to accommodate larger K values in the future through a sparse table implementation.

Baseline. Besides comparing to the baseline novelty model without user alignment, we also compare the proposed method to existing production models: (1) **Exploration-oriented** models include: *Hierarchical contextual bandit* (Song et al., 2022) obtain the next clusters through a tree-based LinUCB; *Neural linear bandit*-based DNN model (Su et al., 2024) to predict the next novel cluster. Although these models are tailored to explore user interests, they are trained on interest transitions ex-

isting in the system and therefore are still subject to the feedback loop. (2) **Exploitation-oriented** models include a regular *two-tower* model (Yang et al., 2020) and *transformer-based* (Chen et al., 2019; Shaw et al., 2018) sequential model trained on all positive user feedback. Our live experimental results demonstrate our proposed method can lead to recommendation which are more novel and of better quality compared to these existing models.

5.2 Model Finetuning and Offline Evaluation

We used offline metrics to guide model training, checkpoint selection, and hyper-parameter searching (e.g. score normalization strategy). Offline evaluation is done on a holdout set of interest cluster sequences, the novel interest transitions and user’s feedback scores. During evaluation, the alignment reward model scores and ranks the interest cluster transitions for each input cluster sequence. We compare this model-generated ranking against the ground-truth ranking from the user feedback. Performance is measured using F1@K (i.e., the harmonic mean of precision and recall), and NDCG@K metrics, with K being the number of interest clusters served in live traffic.

As shown in Figure 3, the offline metrics improve consistently over a random baseline throughout the alignment model’s training process. These results underscore the importance of incorporating user feedback alignment into our inference scaling approach. Furthermore, the offline evaluation guided the hyper-parameter tuning, allowing us to optimize the reward model’s performance and prevent overfitting. In live A/B experiments, we deployed two arms: one favorable arm using an alignment model trained for 50,000 steps (where F1 converged in offline evaluation as shown in Figure 3), and another arm using an alignment model trained for 100,000 steps beyond the favorable converging point as comparison. We observed significantly improved user satisfaction with the favorable arm, as evidenced by a larger positive playback rate (PPB) gain – indicating better alignment with user preferences. This finding is consistent with our hypothesis that extensive training beyond the convergence point can lead to overfitting. While NDCG encourages the model to reproduce the exact ranking from user feedback, F1@K focuses on the model’s ability to identify the top- K most relevant clusters, which is more crucial for our top- n selection task. Memorizing the exact rankings is unnecessary and potentially detrimental to the ex-

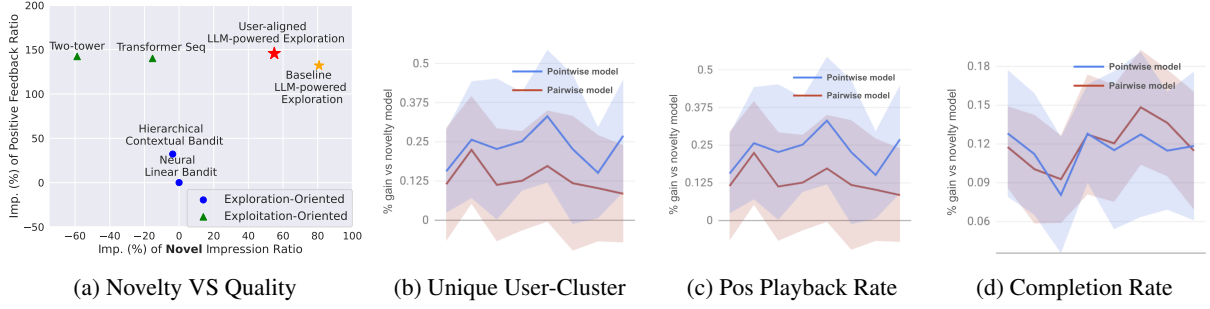


Figure 4: (a) The proposed method still recommends the highest percentage of novelty compared to the rest of the system. (b)(c)(d) Compared to the novelty model baseline, the alignment model further expands users’ interest with higher user satisfaction.

ploration of novel and engaging recommendations.

5.3 Results and Analysis

Novelty and Quality. In Figure 4 (a), we compare the proposed method with various baseline models currently in production. Using the performance of Hierarchical contextual bandit (Song et al., 2022) as the base, we measure improvement of novelty and quality of other models in our system. Specifically, we plot the increase in the novel impression ratio (impressions from interest clusters the user has never interacted with) to highlight recommendation novelty (x-axis), and the increase in positive playback rate to demonstrate recommendation quality (y-axis). We observed that aligning the novelty-focused novelty model with user preference results in higher users’ positive playback ratio at a slight cost of novelty. Nonetheless, the proposed method still has the highest novel impression ratio compared to the rest of the system. Additionally, our method achieves significantly better quality than existing exploration-oriented methods, even surpassing the exploitation-oriented methods. It is rare in the recommendation systems to achieve high novelty and user satisfaction simultaneously. This means through user feedback alignment, we moved our model to a more optimal point in the operation curve – over user satisfaction and engagement improved while the novelty is still the highest in the system.

Increased User Satisfaction. In Figure 4(c), (d), the x-axis represents the experiment periods (the exact dates are redacted), and the y-axis shows the relative percentage difference between the experiment and control. We observed an increase in the positive playback rate and the completion rate of the recommended content, indicating an increased user satisfaction on the platform.

User Interest Exploration. To measure if the recommenders encourage users to explore new inter-

ests, we use unique engaged user-cluster (UEUC), which tracks the number of unique user-cluster engagement pairs. Figure 4(b) shows that our proposed user feedback alignment method not only improves the user satisfaction but also improves the number of user interests. This means our method improves the exploration efficiency. We also notice UEUC is higher for more active users, potentially because the reward model aligns more closely with the preferences of core users who contribute a larger portion of the user feedback training data.

Pairwise vs Pointwise Label. The live experiment results shown in Figures 4(b), (c), and (d) demonstrate a performance comparison between alignment models trained with pairwise labels and those trained with pointwise labels. Both models positively impact user’s interest size and satisfaction, with the pointwise model slightly outperforming the pairwise model. This indicates normalizing users’ feedback per the feedback’s prior helps. Pairwise model learns the relative rank of the novel clusters and its scoring of new cluster may be uncalibrated, thus negatively impacting the performance. We also observed that the pointwise model training is 2x faster. Hence the pointwise model was deployed to production.

6 Conclusion

In this paper, we leverage the hierarchical planning paradigm of LLMs for large-scale recommender systems, separating the exploration and exploitation goals into two distinct models during high-level planning. We share our successful approach to improving alignment using collective user feedback gathered from LLM-powered recommendation systems. Live experiments on a large-scale recommendation platform demonstrate that our proposed method enhances exploration efficiency while simultaneously increasing user engagement.

References

- Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *RecSys*.
- Bo Chang, Changping Meng, He Ma, Shuo Chang, Yang Gu, Yajun Peng, Jingchen Feng, Yaping Zhang, Shuchao Bi, Ed H Chi, et al. 2024. Cluster anchor regularization to alleviate popularity bias in recommender systems. In *Companion Proceedings of the ACM Web Conference 2024*, pages 151–160.
- Minmin Chen. 2021. Exploration in recommender systems. In *RecSys*.
- Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a reinforce recommender system. In *WSDM*.
- Minmin Chen, Yuyan Wang, Can Xu, Ya Le, Mohit Sharma, Lee Richardson, Su-Lin Wu, and Ed Chi. 2021. Values of user exploration in recommender systems. In *RecSys*.
- Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucuri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498*.
- Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt’s capabilities in recommender systems.
- Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chatrec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.
- Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. Vip5: Towards multimodal foundation models for recommendation.
- Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems.
- Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. Gpt4rec: A generative framework for personalized recommendation and user interests interpretation.
- Jianghao Lin, Rong Shan, Chenxu Zhu, Kounianhua Du, Bo Chen, Shigang Quan, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024a. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation.
- Jianghao Lin, Rong Shan, Chenxu Zhu, Kounianhua Du, Bo Chen, Shigang Quan, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024b. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. In *Proceedings of the ACM on Web Conference 2024*, pages 3497–3508.
- Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023a. Is chatgpt a good recommender? a preliminary study.
- Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023b. A first look at llm-powered generative news recommendation.
- Khushhall Chandra Mahajan, Amey Porobo Dharwadker, Romil Shah, Simeng Qu, Gaurav Bang, and Brad Schumitsch. 2023. Pie: Personalized interest exploration for large-scale recommender systems. In *Companion Proceedings of the ACM Web Conference 2023*, pages 508–512.
- Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *CIKM*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*, pages 3464–3475.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.
- Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, et al. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 1039–1044.
- Yu Song, Shuai Sun, Jianxun Lian, Hong Huang, Yu Li, Hai Jin, and Xing Xie. 2022. Show me the whole world: Towards entire item space exploration for interactive personalized recommendations. In *WSDM*.
- Yi Su, Xiangyu Wang, Elaine Ya Le, Liang Liu, Yuening Li, Haokai Lu, Benjamin Lipshitz, Sriraj Badam, Lukasz Heldt, Shuchao Bi, et al. 2024. Long-term

value of exploration: Measurements, findings and algorithms. In *WSDM*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Jianling Wang, Haokai Lu, James Caverlee, Ed Chi, and Minmin Chen. 2024a. Large language models as data augmenters for cold-start item recommendation. *arXiv preprint arXiv:2402.11724*.

Jianling Wang, Haokai Lu, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Ningren Han, Shuchao Bi, Lexi Baugher, et al. 2024b. Llms for user interest exploration in large-scale recommendation systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 872–877.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 12–22.

Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H Chi. 2020. Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion Proceedings of the Web Conference 2020*.