

LLMs for User Interest Exploration in Large-scale Recommendation Systems

Jianling Wang*
Google DeepMind
jianlingw@google.com

He Ma
Google
htm@google.com

Shuzhou Zhang
Google
shuzhouz@google.com

Lexi Baugher
Google
baugher@google.com

Haokai Lu*
Google DeepMind
haokai@google.com

Yueqi Wang
Google
yueqi@w@google.com

Ningren Han
Google
peterhan@google.com

Ed H. Chi
Google DeepMind
edchi@google.com

Yifan Liu
Google
yifanliu@google.com

Yang Gu
Google
greeness@google.com

Shuchao Bi
Google
shuchaobi@google.com

Minmin Chen
Google DeepMind
minminc@google.com

ABSTRACT

Traditional recommendation systems are subject to a strong feedback loop by learning from and reinforcing past user-item interactions, which in turn limits the discovery of novel user interests. To address this, we introduce a hybrid hierarchical framework combining Large Language Models (LLMs) and classic recommendation models for user interest exploration. The framework controls the interfacing between the LLMs and the classic recommendation models through “interest clusters”, the granularity of which can be explicitly determined by algorithm designers. It recommends the next novel interests by first representing “interest clusters” using language, and employs a fine-tuned LLM to generate novel interest descriptions that are strictly within these predefined clusters. At the low level, it grounds these generated interests to an item-level policy by restricting classic recommendation models, in this case a transformer-based sequence recommender to return items that fall within the novel clusters generated at the high level. We showcase the efficacy of this approach on an industrial-scale commercial platform serving billions of users. Live experiments show a significant increase in both exploration of novel interests and overall user enjoyment of the platform.

CCS CONCEPTS

• Information systems → Information retrieval.

* indicates Equal Contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

KEYWORDS

Large Language Models, Recommendation System, User Interest Exploration

ACM Reference Format:

Jianling Wang*, Haokai Lu*, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Ningren Han, Shuchao Bi, Lexi Baugher, Ed H. Chi, and Minmin Chen. 2018. LLMs for User Interest Exploration in Large-scale Recommendation Systems. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Recommendation systems are indispensable in helping users navigate the vast and ever-growing content on the web nowadays. These systems however are often subject to a strong feedback loop [5, 26], recommending items similar to a user’s past behavior. Classic recommendation systems infer a user’s next interest based on their historical interactions. While this can be effective for short-term engagement, it limits users from discovering novel interests, leading to content fatigue. Recent research highlights the importance of **user interest exploration** [7, 9, 25, 28, 29], aiming to introduce diverse content that goes beyond a user’s historical preferences. Effectively introducing novel interests to users are however challenging due to the vast interest space and the high uncertainty of a user’s affinity to previously unseen interests [9, 32].

Recent breakthroughs in Large Language Models (LLMs) [1, 4, 30] and other foundation models offer exciting opportunities to revolutionize recommendation systems [2, 10, 11, 22, 31]. The pre-trained world knowledge in these models holds the potential to break recommendation feedback loops by introducing diverse and serendipitous recommendations, addressing the challenge of user interest exploration. While prior work [13, 17, 20, 21] has demonstrated the potential of using LLMs for recommendation by translating recommendation problems into natural language processing tasks, deploying these approaches in *real-world* industrial recommendation systems remain extremely challenging as: (1) unlike

domain-specific recommendation models, LLMs lack deep knowledge of the massive, and rapidly evolving item corpus on industrial-scale online platforms (e.g., more than 500 hours of content are uploaded to YouTube every minute [3, 15], a new track is uploaded to Spotify every second [18]); (2) off-the-shelf LLMs are unaware of the collaborative signals from users, failing to capture domain-specific user behaviors; and (3) the latency and cost of serving LLMs per user request are prohibitively large, cannot meet the $O(100\text{ms})$ response time expected and production Query-Per-Second (QPS) required on industrial recommendation platforms.

To overcome the above challenges, we introduce a **hybrid hierarchical planning** paradigm combining LLMs and classic recommendation models (as shown in Figure 1) for user interest exploration in large-scale recommendation systems. At the high level, considering the massive number of incoming items in the system, instead of directly predicting the next item, we use LLMs to infer the next novel interest. At the low level, to leverage classic recommendation models with strong personalization, we ground these novel interests to item recommendations by "restricting" conventional transformer-based sequence models [8, 27] to items within the "clusters" defined by those novel interests. By combining the best of both worlds, the hybrid approach leverages LLMs' reasoning and generalization capability in exploring user's novel interests effectively, at the same time bridges the knowledge gap by relying on domain-specific models for actual item recommendation. We further perform supervised fine-tuning (SFT) with real-world novel consumption behaviors for in-domain user alignment, and also enable LLMs to perform **controlled generation**, producing novel interest descriptions that directly match one of the pre-defined clusters. The controlled generation allow algorithm designers to easily define the granularity of the interests generated by LLMs for different applications, which is critical for effectively exploring the interest space. We find diversification and label balance treatment while curating the SFT data significantly mitigate the long-tailed distribution of LLM generation, and thus improves interest exploration efficiency. To address the LLM inference challenge, we propose to use topically coherent interest clusters with cluster-level descriptions to represent recommendation objects, i.e., both the historical user interests and the recommended next novel interests. By using a small number of historical consumed clusters as high-level user interest, we pre-compute the novel interest transitions offline with LLM bulk inference, which can be then be served online with simple table lookup operations. In summary, we make the following contributions:

- We propose a hybrid hierarchical framework that combines LLM's reasoning and generalization capabilities with classic recommendation models with strong personalization and grounded item corpus knowledge for effective user interest exploration.
- We fine-tune LLMs using a diverse and balanced set of novel interest transitions curated from real-world user interactions for controlled generation and user behavior alignment, to ensure LLMs generate novel interests that match one of the predefined interest "clusters" and align with actual user behaviors.
- We propose to adopt topical clusters instead of items to represent user's high-level interests. The coarser representation allows us to limit the length of historical cluster sequence used

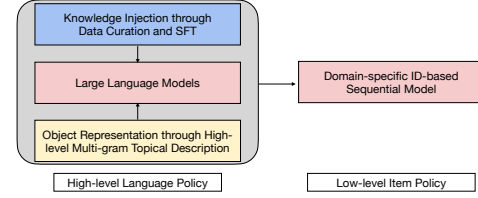


Figure 1: LLM-powered hybrid hierarchical planning diagram for user interest exploration.

to represent dynamic user interests and move the expensive LLM inference to offline stage, making it feasible to serve LLM generated novel interest transitions online.

- We validate our method through live experiments on a large commercial recommendation platform with billions of users. The results clearly show our approach successfully expands user interests while boosting user enjoyment of the platform demonstrated through more active users with longer dwell time.

2 RELATED WORK

LLMs for Recommendation Systems. Application of LLMs to recommendation systems is a rapidly growing research area. Some studies explore using LLMs directly for generating recommendations [2, 11, 14, 17, 21, 23], while others focus on augmenting traditional recommendation models with LLM-powered feature engineering [16, 35] or enriched user/item representations [19, 24, 33]. The computational cost of LLMs however presents a critical challenge. Directly using them for large-scale retrieval is expensive and hinders adoption. Wang et al. [31] use LLMs as data augmenters for conventional recommendation systems during training, to improve model performance without additional serving cost. Different from prior work, we focus on directly incorporating LLM-generated content to break the feedback loop, aiming for more diverse and serendipitous recommendations while maintaining efficiency.

User Interest Exploration. Prior research has established the benefits of User Interest Exploration in recommendation systems, demonstrating its ability to expand user preferences and enhance long-term engagement [7, 9, 29]. However, a key challenge lies in the inherent closed-loop nature of existing systems [5, 9, 26]. Training data is primarily derived from past user-item interactions, limiting the system's ability to explore truly novel interests. While methods like PIE [25] address this to certain extent through user-creator affinity and online bandit formulations, they remain confined by the system's internal knowledge [9]. Our work introduces a novel approach that integrates world knowledge from LLMs to overcome these limitations.

3 METHOD

In this section, we introduce the hybrid hierarchical planning paradigm and the LLM fine-tuning process designed to enable controlled generation and user behavior alignment, to apply LLMs in real-world large-scale recommenders.

Prompt

The short-form videos I watched most recently are in the following clusters:

Cluster 1: {Bus, Truck, Road, Commercial vehicle, Bus Simulator Indonesia, Om Telolet Om, Accident, Motor vehicle, Hino Motors, Tour bus service}

Cluster 2: {Coconut, Banana, Fruit, Durian, Fruit, Oil palms, Mango, Jackfruit, Durian, Papaya}.

Each cluster is described with salient phrases or entity names. With less than 30 words, generate a new and different short-form video cluster I will watch next with highly specific salient phrases or entity names, with a prefix that says "New cluster: "

Label

New cluster: Sharks, Underwater environment, Underwater diving, Dolphin, Orca, Mermaid, Penguins, Whale, Fish, Ocean

Figure 2: Prompt for Novel Interest Prediction when $K = 2$.

3.1 Preliminaries

The sheer number of items and the constant influx of new items on online platforms make LLM planning at the individual item-level infeasible. Instead, we leverage the planning capabilities of LLMs at the item interest-level to reduce the planning space. A prerequisite for efficient hierarchical planning is a set of high-quality *item interest clusters*, where items within each cluster are topically coherent. Following the same procedure as in [6], we group items into N traffic-weighted equal sized **clusters** based on their topical coherence, a method proven to scale well to the magnitude of our problem. To create these clusters, we first represent each item as a 256-dimensional embedding based on its metadata (title, hashtags, etc.) and content (frames and audio). Then, we connect items in a graph based on their similarity and cluster it into traffic-balanced clusters. This clustering process is repeated multiple times to create a 4-level tree structure, with each item associated with different tree levels. Higher-level clusters represent broader topics, while lower-level clusters represent more specific ones. These clusters in each level, denoted by $C^l = \{c_1^l, c_2^l, \dots, c_{M_l}^l\}$, $l = 1, 2, 3, 4$, represent different user interests, with each cluster linked to a **set of keywords** describing its theme. Here M_l denotes the number of clusters within level l , with $M_4 > M_3 > M_2 > M_1$. Each item belongs to a single interest cluster in each level. As discussed in [6], we focus on level-2 clusters to balance granularity and feasible planning space¹.

3.2 Hybrid Hierarchical Planning

The hybrid approach combines a LLM to produce a language policy that generates novel interests on the high-level, and a classic recommendation models to produce an item policy that grounds these language-based interests to the low-level item space. Such a hybrid approach combines the strength of LLMs in reasoning and generalization, and domain-specific recommendation models in handling item dynamics and enhanced personalization.

High-level language policy. Given the historical user interest representation by language, we first use LLM to learn a high-level language policy that generates novel user interests. Instead of using item descriptions to represent users, we propose to adopt cluster descriptions (i.e., a set of keywords) to represent user’s consumption history, i.e., a user’s historical interest is represented as a sequence of her K most recent interacted unique clusters, with each cluster represented by its description. Specifically, with a user’s previously consumed unique clusters, we can ask LLM to generate the next novel interest with the prompt illustrated in Figure 2.

¹To reduce notation complexity, we drop level l in cluster moving forward.

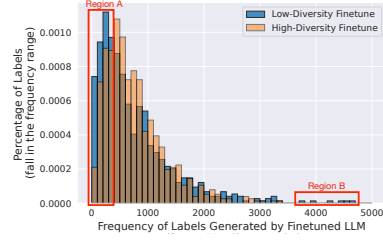


Figure 3: Label (i.e., generated by fine-tuned LLM) Distribution: X-axis represents label frequency; Y-axis represents the percentage of labels within each frequency range.

Practical implication. One major challenge of deploying LLM to industrial-scale recommendation system lies in its prohibitively high inference cost failing to meet latency and QPS requirements. Empirically, we find that relying on a small number of historical clusters to represent each user (e.g., $K = 2$) can effectively balance representation granularity and computation efficiency. In our experiment, the level-2 clustering produced 761 clusters. We can therefore enumerate all $761 * 761 = 579,121$ cluster pairs and perform a batch inference with LLM to obtain novel interests for each cluster pair under just a few hours. These novel interests, together with the input cluster pairs, could be stored in a table. During *online serving* as a new user request comes in, we first represent the user by sampling $K = 2$ items from their watch history², and convert them into the cluster pair for lookup to determine the recommended novel interest cluster.

Low-level item policy. Once the language based novel user interest is obtained, the next step is to convert it to item-level recommendation policy. A straight-forward approach is to rely on search engine [21] to retrieve the most relevant items according to the keywords of the novel interest. The search results however often lack personalization since these language based novel interests can still be broad and lack specificity. To enhance personalization, we propose to reuse domain specific recommendation models, specifically *transformer-based sequential recommender* model [8, 27], but restrict the items to clusters prescribed by the language based novel interests. Specifically, we follow the two steps: (i) map the generated novel interests to cluster ID space, and (ii) restrict the original item level softmax policy on these cluster IDs, to retrieve items **only** from these clusters.

Controlled generation. The remaining challenges are how to specify the granularity of the LLM generated interests and map the generated novel interests to cluster IDs. Free-form responses from LLMs can be arbitrary, which are unlikely to be directly matched to the predefined cluster descriptions. We control the granularity of the generation through the hierarchical clustering and picking the cluster level explained in Section 3.1. Furthermore, appropriate finetuning as detailed in Section 3.3 enables LLMs to speak the language of interest clusters, producing cluster descriptions that exactly match one of the predefined clusters.

²30-day user history are used in sampling, and items with high-quality interactions are more likely to be sampled.

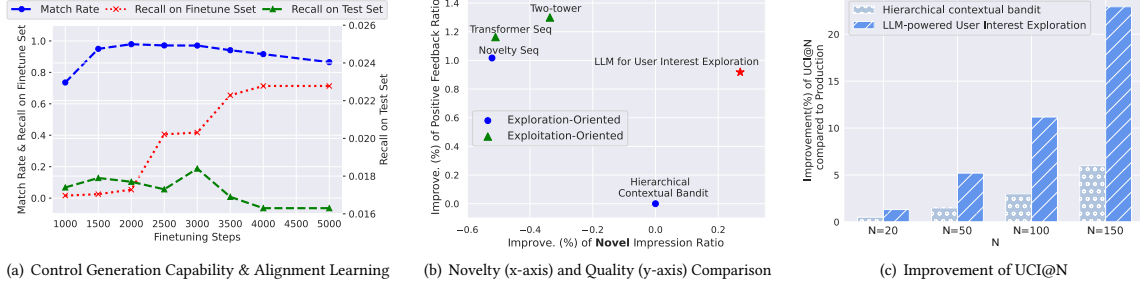


Figure 4: (a) Model Finetuning Process. (b) and (c) Comparison between different recommenders in live experiments.

3.3 Fine-Tuning for User Behavior Alignment

LLMs trained on massive publicly available data on the Internet contains rich global knowledge, it however lacks capabilities to perform: 1) controlled generation (i.e., generate in the interest cluster space) and 2) domain-specific user behavior alignment. We propose to inject these domain specific knowledge through supervised fine-tuning using a dataset curated from real user watch histories on the commercial platform. The quality of data used for fine-tuning is thus crucial for its success.

Diversified Data Curation. Take $K=2$ as an example. Each fine-tuning data sample, denoted as $[(C_1, C_2), C_L]$, consists of a cluster pair (C_1, C_2) to form the prompt and the subsequent novel cluster C_L as the finetune label. By definition of novelty, C_L must be different from C_1 and C_2 . Initially, we gathered approximately 250K $[(C_1, C_2), C_L]$ data samples from our log, focusing solely on high quality interactions³. These samples are then grouped by their labels, and the top-10 most frequently occurring cluster pairs are selected for each label, forming the final data samples which is diverse and cover all the labels. Following these steps, we obtain $761 * 10 = 7,610$ data samples (10 per label cluster) and proceed to perform supervised fine-tuning for the LLM using these samples.

In Figure 3, we plot the distributions of interest clusters generated by fine-tuned LLMs on the 579,121 context cluster pairs. When using finetuning data of low diversity, where we randomly select 7,610 transitions and their corresponding subsequent clusters from the initial 250K to form the data, the fine-tuned LLM generates interests that are highly-skewed, where a few generated clusters have very high frequencies (depicted as Region B). When we increase the diversity of our fine-tuning data, these dominant labels disappear, and the number of generated clusters with very low frequency (depicted in Region A) also decreases. Ensuring the fine-tuning data covers all clusters evenly allows us to address the long-tailed distribution issue in the model’s generated clusters. This treatment not only mitigates the feedback loop effect in behavior data but also enhances overall user satisfaction as shown in Section 4.

Control Generation Capability & User Behavior Alignment. The number of fine-tuning steps determines the balance between the LLM’s global and task-specific knowledge. Our fine-tuning process has two main objectives: (1) **controlling** LLM generation to speak the language of interest clusters. We evaluate the *match rate* of the generation from the fine-tuned LLM to determine if the output matches exactly with one of the cluster descriptions; and (2) **aligning** with real-world user transitions, measured by comparing

³In other words, these samples are novel interest transitions existed in the log where a user was successfully introduced a novel interest.

the fine-tuned LLM’s output with the successful user interest transition in both fine-tuning and test set to compute *recall*. A higher recall indicates LLM learning the domain specific novel transitions from the fine-tuned data, and aligning with user behaviors.

In Figure 4 (a), with batch size of 16, we illustrate the changes in match rate and recall as the fine-tuning steps progress. We note that formatting learning, i.e., learning the language of interest clusters, kicks in first, peaking at around 2,000 steps. With a high match rate (over 99%), we can efficiently map the generation to cluster ID space and restrict the original item-level softmax policy on these clusters. Subsequently, the model begins to align with user behaviors, resulting in a significant increase in recall (on the fine-tuned set). Moreover, we find that the recall for a separate test set increases following transition alignment, reaching its peak at around 3,000 steps before gradually decreasing. Therefore, we select models fine-tuned with 3,000 steps. Note that the recall on the test set is much lower than that of the fine-tuning set, indicating LLM is still relying heavily on its global knowledge instead of memorizing interest transitions in the log while generating novel interests.

4 LIVE EXPERIMENTS

4.1 Experimental Setup

We conduct a series of live experiments on a *commercial short-form video recommendation platform that serves billions of users*. Our experiments are conducted with **Gemini-Pro** [12], but the same fine-tuning process and pipeline can be readily adapted to other LLMs. We set the number of historical clusters for LLM inference $K = 2$, it however can easily scale to accommodate larger numbers in future iterations with a sparse table.

Baseline. We compare the proposed method to existing production models: (1) **Exploration-oriented** models include: a *Novelty-enhanced sequence recommender* [9] trained with labels from both positive and novel items whose clusters have not appeared in user’s consumption history before; *Hierarchical contextual bandit* [28] based on the hierarchical clusters introduced in 3.1 to explore user’s interests through a tree-based LinUCB to obtain the next clusters, from which the sequential model is then used to restrict the retrieval to items. Although these models are tailored to explore user interests, they are trained on interest transitions existing in the system and therefore are still subject to the feedback loop. (2) **Exploitation-oriented** models include a regular *two-tower* model [34] and *transformer-based* [8, 27] 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 in better quality compared to these existing models.

4.2 Results and Analysis

Novelty and Quality. In Figure 4 (b), we compare the proposed method with various baseline models currently in production. Using the performance of Hierarchical contextual bandit [28] as the reference, we measure the improvement of the other models. Specifically, we plot the increase in ratio of novel impressions (considering only impressions from interest clusters the user has never interacted with) to highlight recommendation novelty (x-axis), and the increase in positive feedback rate to demonstrate recommendation quality (y-axis). The proposed method recommends more novel items compared to all the baseline methods (to the right on x-axis). Additionally, it achieves much better quality than existing exploration-oriented methods, comparable to exploitation-oriented methods (high on x-axis). In other words, the proposed method presents an effective approach to introduce users to novel interests that are of interests to the user.

User Interest Exploration. To measure if the recommenders encourage users to explore new interests, we use a metric called $UCI@N$, which tracks the number of users who have consumed items from N unique clustered interests within the past 7 days. Higher $UCI@N$ indicates more users are consuming N interests. By monitoring $UCI@N$ for different values of N (20 to 200), we can gauge the effectiveness of our system for user interest exploration. Figure 4(c) summarizes the improvement of our method compared to Hierarchical Contextual Bandit, to evaluate its effectiveness in user interest exploration. Notably, our proposed method shows very significant improvement compared to the prominent exploratory model currently deployed in production for different values of N .

User Growth. At the same time, we monitor increase in overall watch time and number of active users who had total watch time ≥ 10 minute (in Figure 5), to measure user growth on the short-form video platform. 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, which excludes the proposed system. Our method successfully broadens user interest by recommending diverse and novel content, with user growth. This underscores the quality and relevance of the recommended novel content.

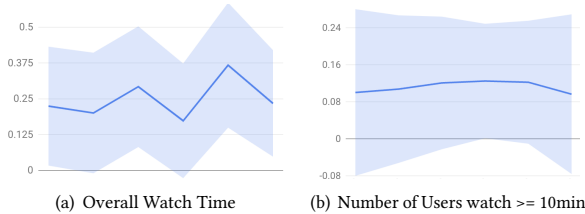


Figure 5: The proposed method drives user growth.

5 CONCLUSION

We present a hybrid approach to leverage LLMs for user interest exploration. It combines the strength of LLMs in reasoning and generalization, and the grounding of classic recommendation models. We showcase a successful recipe to inject domain-specific recommendation knowledge to LLMs for controlled generation and user behavior alignment. Extensive testing on a commercial platform with billions of users has yielded significant improvements in both

exploration of novel interests and user growth. The future work will focus on taking long-term effects into account to further improve hierarchical planning with LLMs for recommendation systems.

REFERENCES

- [1] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- [2] 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. *arXiv preprint arXiv:2305.00447*.
- [3] Youtube Official Blog. 2023. *YouTube by the Number*. Retrieved January, 2023 from <https://blog.youtube/press/>
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.
- [5] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *RecSys*.
- [6] 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 on Web Conference 2024*. 151–160.
- [7] Minmin Chen. 2021. Exploration in recommender systems. In *RecSys*.
- [8] 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*.
- [9] 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*.
- [10] Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Churci, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498* (2023).
- [11] 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. *arXiv preprint arXiv:2305.02182*.
- [12] Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [13] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *ResSys*.
- [14] Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. *arXiv preprint arXiv:2305.14302*.
- [15] James Hale. 2019. *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*. Retrieved January, 2023 from <https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/>
- [16] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *TheWebConf*.
- [17] 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. *ECIR*.
- [18] Tim Ingham. 2023. *Over 60,000 Tracks are Now Uploaded to Spotify Every Day. That’s Nearly One per Second*. Retrieved January, 2023 from <https://www.musicbusinessworldwide.com/over-60000-tracks-are-now-uploaded-to-spotify-daily-thats-nearly-one-per-second/>
- [19] Chen Li, Yixiao Ge, Jiayong Mao, Dian Li, and Ying Shan. 2023. TagGPT: Large Language Models are Zero-shot Multimodal Taggers. *arXiv preprint arXiv:2304.03022*.
- [20] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text Is All You Need: Learning Language Representations for Sequential Recommendation. In *KDD*.
- [21] 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. *arXiv preprint arXiv:2304.03879*.
- [22] Jianghao Lin, Rong Shan, Chenxu Zhu, Kounianhua Du, Bo Chen, Shigang Quan, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. *TheWebConf*.
- [23] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*.
- [24] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023. A First Look at LLM-Powered Generative News Recommendation. *arXiv preprint arXiv:2305.06566*.

- [25] 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*. 508–512.
- [26] Masoud Mansoury, Himan Abdollahpour, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *CIKM*.
- [27] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* (2018).
- [28] 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*.
- [29] 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*.
- [30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [31] Jianling Wang, Haokai Lu, James Caverlee, Ed Chi, and Minmin Chen. 2024. Large Language Models as Data Augmenters for Cold-Start Item Recommendation. *arXiv preprint arXiv:2402.11724* (2024).
- [32] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. 2024. Result Diversification in Search and Recommendation: A Survey. *TKDE* (2024).
- [33] Yunjia Xi, Weiwen Liu, Jianghao Lin, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. *arXiv preprint arXiv:2306.10933*.
- [34] 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*.
- [35] Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, and Qi Liu. 2022. Tiny-newsrec: Effective and efficient plm-based news recommendation. In *EMNLP*.