Adaptive and Representative Multi-Interest Modeling for Recommendation with Large Language Model

Ziyan Wang[1], Yingpeng Du[1]*, Tianjun Wei[1], Haoyan Chua[1], Jieyi Bi[1], Jie Zhang[1], Zhu Sun[2]

[1] College of Computing and Data Science, Nanyang Technological University
[2] Information Systems Technology and Design Pillar, Singapore University of Technology and Design
{wang1753, haoyan001, jieyi001}@e.ntu.edu.sg
{yingpeng.du, tianjun.wei, zhangj}@ntu.edu.sg
zhu_sun@sutd.edu.sg

Abstract

Large language models (LLMs) show potential for multi-interest analysis of users in recommender systems, going beyond heuristic assumptions in existing methods, e.g., co-occurring items indicate the same interest. Despite the effectiveness, two key challenges remain. First, the granularity of raw generation of LLMs for multi-interests is agnostic, possibly leading to overly fine or coarse interest grouping. Second, adopting LLM to analyze individual user behaviors lacks a global perspective on how items relate across users. In this paper, we propose an LLM-driven adaptive and representative multi-interest modeling framework to address the challenges. At the user-individual level, we exploit LLM analysis and alleviate the agnostic granularity by adaptively aggregating semantic clusters to collaborative multi-interests. At the user-crowd level, to mitigate the limited insights in individual behaviors, we formulate a max covering problem to expand the scope of LLM analysis with compactness and representativeness, disentangling interest representations from global perspectives. Experiments on real-world datasets show that our approach outperforms various baselines.

1 Introduction

Recommender systems (RSs) play a crucial role in personalized user experience, while modeling users' multi-faceted and dynamic interests remains challenging. To bridge this gap, multi-interest recommendation methods (Li et al., 2019; Zhang et al., 2022) use multiple representations for each user to capture users' diversified interest facets behind their behaviors.

Despite their effectiveness, existing methods often rely on heuristic assumptions, such as similar items indicate the same interest for users, where similarity can be measured through item embeddings (Ma et al., 2020), co-occurrence statistics (Du et al., 2024b), or auxiliary information (Chai et al., 2022). However, due to the sparsity in user-item interactions as well as the incompleteness in auxiliary data, accurately measuring item similarity and user interests remains a significant challenge. As in Figure 1(a), though the items within the red box belong to the 'Accessories' attribute, they significantly differ in functionalities and appeal to different user interests. To address these problems, we leverage large language models (LLMs) with their rich knowledge and powerful reasoning capabilities (Wu et al., 2024), analyzing users' semantic multi-interests beyond their interaction records and auxiliary information. LLMs can offer semantic guidance for multi-interest extraction, providing more accurate guidance on multi-interest analysis as illustrated in blue boxes.

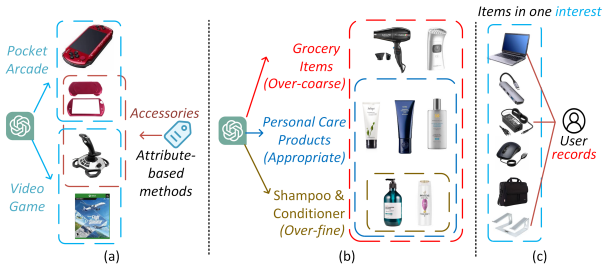

Figure 1: (a) Difference between attribute- and LLM-driven multi-interest analysis. (b) LLM-driven multi-interest analysis leads to varying granularity. (c) Individual users' behaviors lack global item relationships.

Although leveraging LLMs offers a promising way for multi-interest modeling, directly using them as a black-box is not a one-size-fits-all solution, as there remain two significant challenges. **First**, the granularity of LLM-driven multi-interest analysis is agnostic, i.e., overly-fine or overly-coarse item division among users' behaviors, making it hard to model their multi-interests accurately. As in Figure 1(b), LLMs might categorize interests with overly-fine distinctions among items

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 2484–2496

that should belong to the same interest ('Shampoo & Conditioner' in brown), or with overly-coarse groupings that fail to capture interest discrimination among items ('Grocery Items' in red), while only the 'Personal Care Products' (in blue) is the desired or ideal interest. **Second**, multi-interest analysis for individual users lacks a global perspective on item relations across the entire population. Specifically, relying solely on individual user interactions analyzed by LLMs, which are inherently sparse, may lead to incomplete modeling of multiple interests. As in Figure 1 (c), besides the engaged items in users' behaviors connected by red lines, there remain multiple non-co-occurring items that may also reflect the same user interests.

To address the first challenge, we guide the LLM-driven multi-interest analysis at user-individual level using a tailored prompt to cluster each user's interaction sequence into distinct semantic interest groups. We then introduce an interpretable alignment module that dynamically aggregates LLM-driven semantic clusters and maps them into collaborative interests learned by capsule network. By doing so, we adaptively adjust the granularity to fit user patterns. For the second challenge, we generate synthetic users for LLM-driven analysis at the user-crowd level, ensuring the compactness (limited number of interests with coherent behaviors) and representativeness (covering multiple items). We achieve this by clustering real users with similar preference and solving a max covering problem (MCP) to select synthetic users that span the item space. Finally, we introduce contrastive learning to encourage item concentration within interests and dispersion across interests, enhancing multi-interest representations.

Our key contributions are three-fold. Firstly, we propose a novel LLM-driven Adaptive and Representative Multi-Interest (LARMI) modeling framework to explore semantic information from both the user-individual level and user-crowd level, for more effective multi-interest recommendation. Secondly, we address the issues of LLM's agnostic granularity by designing an adaptive alignment module and MCP optimization with contrastive learning to ensure the compactness and representativeness from a global perspective. Thirdly, we evaluate the proposed method across three real-world datasets to demonstrate its effectiveness in multi-interest modeling.

2 Literature Review

Single- and Multi-Interest Modeling for Recommendation. Single-interest modeling in recommendation include methods built upon recurrent neural networks (Hidasi et al., 2016; Guo et al., 2020), self-attention (Kang and McAuley, 2018; Zhang et al., 2019), transformers (Sun et al., 2019; Xia et al., 2021), and graph convolutional networks (He et al., 2020; Wang et al., 2025), but these methods overlook the diversity in user interests. Current multi-interest modeling methods that adopt capsule networks (Sabour et al., 2017; Li et al., 2019; Xie et al., 2023; Tian et al., 2022) solely rely on users' engaged items. Attention mechanisms are also incorporated (Cen et al., 2020; Xiao et al., 2020). Regularization strategies (Zhang et al., 2022; Lee et al., 2024) stabilize the learning of multiple embeddings. Diffusion model (Le et al., 2025) and item partition objective (Du et al., 2024b) are applied for interest-aware denoising and enhancement. Other methods utilize auxiliary sources such as users' profiles (Chai et al., 2022), timestamps (Chen et al., 2021), items' categories (Liu et al., 2024), and knowledge graphs (Liu et al., 2022).

Large Language Models for Recommendation. LLMs' success has inspired their incorporation in recommendation pipelines (Wu et al., 2026). LLM-as-recommender methods employ LLMs as scoring functions or rankers. Methods fully fine-tune the LLM parameters (Geng et al., 2022; Qu et al., 2024) or conduct parameter-efficient fine-tuning (PEFT) (Bao et al., 2023; Jiang et al., 2025) bridge the gap between LLMs and recommendation tasks. Non-tuning methods align the recommendation objectives for LLMs through zero-shot prompting (Dai et al., 2023; Hou et al., 2024) and in-context learning (Sanner et al., 2023; Bao et al., 2025) strategies. LLM-as-extractor methods apply LLMs for data augmentation. Studies focus on encoding historical behaviors and item attributes produce expressive embeddings (Wang et al., 2024; Harte et al., 2023) to capture complex semantic information. Besides, several methods adopt LLMs to extract additional knowledge such as user profiles (Zheng et al., 2023; Du et al., 2024a), item descriptions (Ren et al., 2024; Wei et al., 2024), and other textual data (Mohbat and Zaki, 2025) through semantic mining.

3 Methodology

This section presents **LARMI**, the proposed LLM-based Adaptive and Representative Multi-Interest

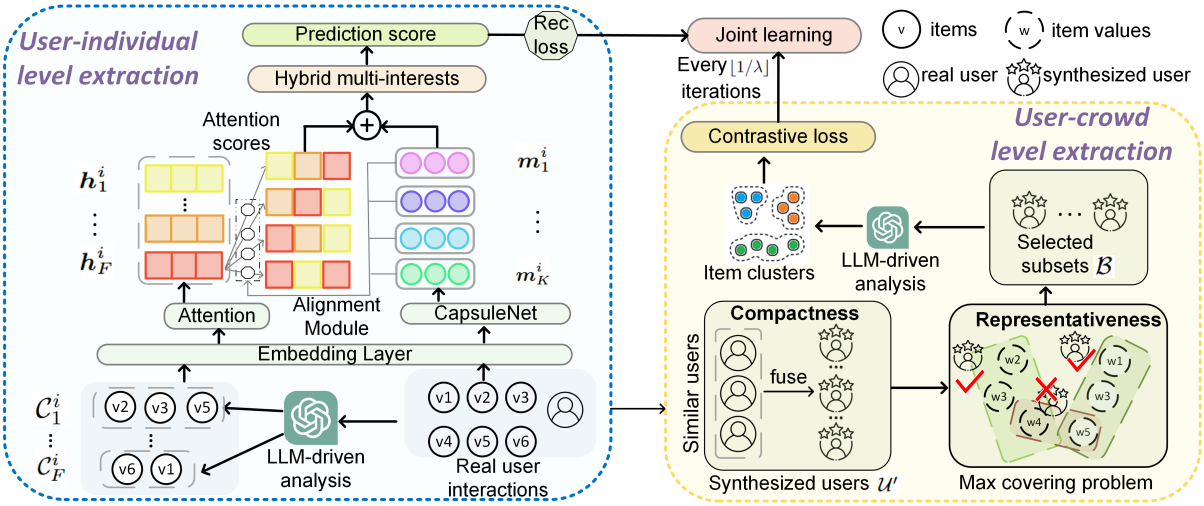

Figure 2: The overall architecture of our proposed LARMI.

modeling framework, which leverages LLM-driven semantics for multi-interest modeling in RSs. We denote the set of M users as $\mathcal{U} = \{u_1, \cdots, u_M\}$ and the set of N items as $\mathcal{V} = \{v_1, \cdots, v_N\}$. Each user $u_i \in \mathcal{U}$ has a sequential behavior series sorted by timestamps denoted as $s_{(u_i)} = \{v_1^i, v_2^i, \cdots, v_L^i\}$, where v_j^i denotes the j-th item engaged by the user u_i and L is the length of the sequence. Besides, we suppose to know item titles in $s_{(u_i)}$ denoted as $t_{(u_i)} = \{t_1^i, t_2^i, \cdots, t_L^i\}$. Given the user's historical behavior, we aim to generate a top-n ranking list containing items that the user is likely to engage in the near future.

3.1 Model Overview

For the user-individual level, we employ the LLM to analyze sequential behaviors for each user, infer distinctive and meaningful semantic multi-interests, and adaptively align with collaborative interests for proper granularity. For the user-crowd level, we synthesize compact and representative users with the MCP optimization, and then bridge the gap between real and synthesized users to expand the LLM analysis scope beyond individual users. Figure 2 shows the overall framework of LARMI.

3.2 User-individual Multi-interest Extraction

We propose to leverage the semantic knowledge of the LLM to guide multi-interest extraction, overcoming the limitation of heuristic assumptions such as co-occurring items implying the same interest of users. Specifically, we prompt the LLM to conduct multi-interest analysis as follows, generating distinctive semantic clusters, each representing a cohesive set of items with shared characteristics.

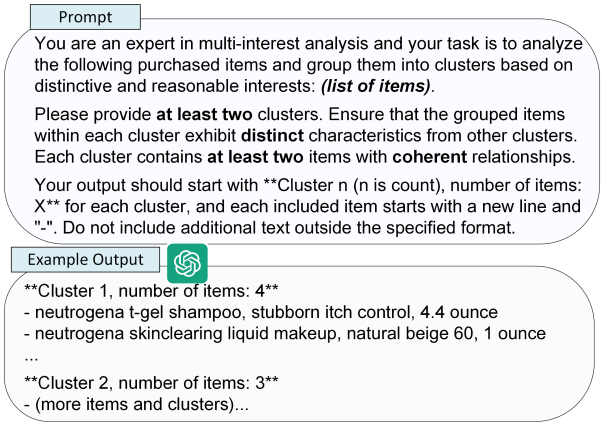

Figure 3: Example output for LLM analysis.

$$\mathcal{C}_1^i, \cdots, \mathcal{C}_F^i = LLM(pmt, t_{(u_i)}), \quad (1)$$

where pmt denotes the multi-interest analysis prompt. $\mathcal{C}_f^i$ denotes the f-th cluster that contains items belonging to the same semantic group by LLM for user u_i. F is an unknown varying number relying on LLMs' analysis and reasoning. We show the prompt and example output in Figure 3.

Intuitively, different clusters reflect distinct aspects (e.g., functionality, preference) for the user. However, the granularity of LLM-driven clusters is agnostic, making it uninterpretable and hard to effectively model multi-interests with overly-fine or overly-coarse clusters. To address this, we propose an adaptive alignment module consisting of an attention mechanism and a projection layer. For over-coarse clusters in LLMs' analysis, we use an attention mechanism to dynamically aggregate the items' representations in a cluster $\mathcal{C}_f^i$, allow-

ing more specific signals to dominate the cluster and sharpen the encoding. The LLM-driven multi-interest representation is defined as:

$$\boldsymbol{h}_f^i = \sum\nolimits_{v_j \in \mathcal{C}_f^i} \alpha_j \cdot \boldsymbol{v}_j, \;\; \alpha_j = \frac{\exp(\mathbf{w}^T \boldsymbol{v}_j + b)}{\sum_{v_k \in \mathcal{C}_f^i} \exp(\mathbf{w}^T \boldsymbol{v}_k + b)}, \tag{2}$$

where $\boldsymbol{v}_j$ denotes item v_j's learned ID embedding, and $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ are learnable weights. Formally, we quantify cluster coarseness as the average pairwise cosine distance of its item embeddings. Higher values indicate over-coarse clusters with diverse items, whereas lower values suggest over-fine clusters of nearly duplicates. Second, the over-fine clusters can be averaged in the previous step, and further merged with collaborative interests based on their similarity. We use capsule network to model the collaborative multi-interests:

$$\boldsymbol{m}_1^i, \cdots, \boldsymbol{m}_K^i = CapsuleNet([\boldsymbol{v}_1^i, \cdots, \boldsymbol{v}_L^i]), \tag{3}$$

where $\boldsymbol{v}_j^i$ denotes the ID embedding of the j-th item engaged by the user u_i, and K is the predefined number of users' multi-interests. Then, we compute attention scores between pairs of semantic and collaborative interest facets for alignment:

$$\boldsymbol{z}_k^i = \sum_{f=1}^{F} \alpha_{kf} \cdot \boldsymbol{h}_f^i, \alpha_{kf} = \frac{\exp\left(\boldsymbol{m}_k^i \cdot \tanh(\boldsymbol{W_1} \boldsymbol{h}_f^i)\right)}{\sum_{f'} \exp\left(\boldsymbol{m}_k^i \cdot \tanh(\boldsymbol{W_1} \boldsymbol{h}_{f'}^i)\right)}, \tag{4}$$

where $\boldsymbol{z}_k^i$ is the aggregation of the semantic clusters related to interests $\boldsymbol{m}_k^i$ of user u_i, α_{kf} is the attention score between LLM-derived semantic clusters $\boldsymbol{h}_f^i$ and collaborative interests $\boldsymbol{m}_k^i$, $\boldsymbol{W_1} \in \mathbb{R}^{d \times d}$ is a learnable projection matrix, and $\tanh(\cdot)$ introduces non-linearity to better capture complex cross-representation relationships.

Therefore, our alignment module allows the cluster granularity to be adaptively adjusted, where specific items can be emphasized through attention weights for sharper interest representation for over-coarse clusters, while similar attention weights can be learned to enable merging over-fine clusters. To take advantage of LLM-driven semantics and collaborative signals, we aggregate them as follows:

$$\boldsymbol{o}_k^i = \boldsymbol{m}_k^i + \boldsymbol{z}_k^i. \tag{5}$$

Thus, we obtain the final hybrid multi-interest representations $\{\boldsymbol{o}_1^i, \cdots, \boldsymbol{o}_K^i\}$ for each user u_i.

3.3 User-crowd Multi-interest Extraction

While user-individual multi-interest modeling leverages LLMs for semantic analysis, it lacks a representative global perspective. Individual users typically interact with only a small subset of items, inevitably preventing all related items from being grouped to the same interest. To address this, we synthesize users with richer behaviors for a more comprehensive LLM analysis. However, simply synthesizing users through random item selection produces dispersed interests and largely increases the scale and cost of LLM inference. To this end, we propose to synthesize users with the consideration of compactness and representativeness principles. **Compactness** ensures that synthesized users have focused interests, with each containing a cohesive set of semantically related items. Otherwise, aggregating unrelated behaviors can lead to fragmented interests and generate sparse, ineffective interest clusters. **Representativeness** maximizes the coverage of unique items across all interests. This avoids redundant LLM analysis and enhances the generalization capability to better represent real-world interest diversity by the synthesized users.

To achieve **compactness** for user synthesis, we formulate a max covering problem (MCP) by grouping cliques of users with overlapping preferences and combine their behaviors to synthesize a user. Specifically, a clique $c_{(u_i')}$ can be generated by clustering similar users w.r.t. a real user u_i, i.e.,

$$c_{(u_i')} = \bigcup\nolimits_{u_g \in \mathcal{N}_{(u_i)}} s_{(u_g)}, \tag{6}$$

where $\mathcal{N}_{(u_i)}$ denotes users who share the most overlapped behaviors to the user u_i. Therefore, each clique can be regarded as the union set of items of a compact synthesized user u_i'. We allow items to belong to multiple interest clusters, enabling LLM-driven analysis to successfully detect distinct intents. However, prompting LLM for all synthesized users leads to redundancy and inefficiency. To this end, we propose selecting representative users for LLM-driven multi-interest analysis. To further achieve **representativeness**, we select a small portion of synthesized users covering as many valuable items as possible across all interests, i.e.,

$$\max_{|\mathcal{B}| \leq Z, \mathcal{B} \subset \mathcal{U}'} \sum\nolimits_{j=1}^{N} \mathbb{I}(v_j \in \bigcup\nolimits_{u_i' \in \mathcal{B}} c_{(u_i')}) \cdot w_{v_j}, \tag{7}$$

where $\mathcal{U}'$ is the set of synthesized users, $\mathcal{B}$ is the selected representative synthesized users with maximal size Z, and w_{v_j} is the value for covering item v_j. Assuming popular items have higher values because they have more impacts, we formulate values and construct a synthesized user-item interaction

matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ based on compactness:

$$w_{v_j} = 1 + \frac{\sum_{i=1}^{M} \mathbb{I}(v_j \in s_{(u_i)})}{\sum_{i=1}^{M} |s_{(u_i)}|}, \quad \boldsymbol{A}_{ij} = \begin{cases} 1 & \text{if } v_j \in c_{(u'_i)}, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where each row indicates the behaviors of a synthesized user based on the corresponding real user. We then formally formulate Equation (7) into a standard MCP. Specifically, we propose to represent the solution $\mathcal{B}$ with an indicator vector $\boldsymbol{x} \in \{0,1\}^M$, where $\boldsymbol{x}_i$ denotes the i-th element of $\boldsymbol{x}$, showing whether the synthesized user u'_i is included in $\mathcal{B}$. Then, Equation (7) can be reformulated as:

$$\begin{aligned} &\max_{|\mathcal{B}| \leq Z, \mathcal{B} \subset \mathcal{U}'} \sum_{j=1}^{N} \mathbb{I}(v_j \in \bigcup_{u'_i \in \mathcal{B}} c_{(u'_i)}) \cdot w_{v_j} \\ \Leftrightarrow &\max_{|\mathcal{B}| \leq Z, \mathcal{B} \subset \mathcal{U}'} \sum_{j=1}^{N} \mathbb{I}\left(\sum_{u'_i \in \mathcal{B}} \mathbb{I}(v_j \in c_{(u'_i)})\right) \cdot w_{v_j} \\ \Leftrightarrow &\max_{|\mathcal{B}| \leq Z, \mathcal{B} \subset \mathcal{U}'} \sum_{j=1}^{N} \mathbb{I}\left(\sum_{u'_i \in \mathcal{B}} \boldsymbol{A}_{ij}\right) \cdot w_{v_j} \\ \Leftrightarrow &\max_{\boldsymbol{x} \in \{0,1\}^M, ||\boldsymbol{x}||_0 \leq Z} \sum_{j=1}^{N} \mathbb{I}\left(\sum_{i=1}^{M} \boldsymbol{x}_i \cdot \boldsymbol{A}_{ij}\right) \cdot w_{v_j}. \end{aligned} \tag{9}$$

To solve the MCP, we adopt a differentiable optimal transport model to collect the compact and representative synthesized users $\mathcal{B}$ following (Wang et al., 2022). Given these synthesized users with rich behaviors, we propose to trigger the LLM to generate distinctive and comprehensive interest clusters for each synthetic user $u'_i \in \mathcal{B}$, i.e.,

$$\mathcal{C'}^{i}_{1}, \cdots, \mathcal{C'}^{i}_{F} = LLM(pmt, t(u'_i)) \tag{10}$$

where $t(u'_i)$ represents the titles of items within the synthesized user u'_i's behaviors $c_{(u'_i)}$, and $\mathcal{C'}^{i}_{f}$ is the f-th cluster generated by the LLM.

As simply producing multi-interest representations of synthesized users has limited impact on real users, we leverage their LLM-driven multi-interests through contrastive learning to refine item distributions, eventually contributing to real users bridged by item representations in a global view. Specifically, we encourage item similarity within the same clusters and dispersion among different clusters in the representation space:

$$\mathcal{L}^{cst}_{u'_i} = -\sum_{v_j \in c_{(u'_i)}} \sum_{v_j * \in \mathcal{C}'(v_j)} \log \frac{e^{\left(\boldsymbol{v}_j^{\intercal} \cdot \boldsymbol{v}_j^* / \tau\right)}}{\sum_{v'_j \in c_{(u'_i)} \setminus \mathcal{C}'(v_j)} e^{\left(\boldsymbol{v}_j^{\intercal} \cdot \boldsymbol{v}'_j / \tau\right)}}, \tag{11}$$

where for each synthesized user u'_i, $\mathcal{C}'(v_j)$ denotes the cluster which contains the item v_j, v_j^* denotes a positive sample that belongs to one same cluster as v_j, and v'_j denotes negative samples that do not occur with v_j. τ controls the sharpness of the similarity distribution. Therefore, for each item in an interest cluster, intra-cluster items are positive instances, and inter-cluster items are hard negative instances. We aggregate the contrastive learning for all synthesized users as the overall loss,

$$\mathcal{L}^{cst} = -\frac{1}{|\mathcal{B}|} \sum_{u'_i \in \mathcal{B}} \mathcal{L}^{cst}_{u'_i}. \tag{12}$$

In summary, we synthesize a compact and representative subset of users with rich behaviors and formulate an MCP optimization for LLM-driven analysis, enabling global perspective for improved multi-interest modeling.

3.4 Multi-task Objective Function

To effectively bridge the adaptive user-individual and representative user-crowd multi-interest modeling, we propose a multi-task learning framework. For real users, we employ hard readout to predict user-item score for recommendation:

$$f(u_i, v_j) = \max_{1 \leq k \leq K} \left(\boldsymbol{o}_k^{i\,\top} \boldsymbol{v}_j \right), \tag{13}$$

where the interest is selected among all interests by the maximum score. For the recommendation task, the objective function can be formulated with the InfoNCE loss as follows:

$$\mathcal{L}^{rec} = -\sum_{(u_i, v_j) \in \mathcal{D}} \log \frac{\exp\left(f(u_i, v_j)\right)}{\sum_{v'_j \in \mathcal{V}} \exp\left(f(u_i, v'_j)\right)}, \tag{14}$$

where $\mathcal{D}$ is the training set for user-item interactions. For selected synthesized users, we employ contrastive learning in Equation (12) to bridge them with real users through item representation learning to enhance multi-interest modeling.

To perform multi-task learning for user-individual and user-crowd multi-interest modeling, our overall objective function aggregates these two goals in a weighted way:

$$\mathcal{L} = \mathcal{L}^{rec} + \lambda \cdot \mathcal{L}^{cst}, \tag{15}$$

where λ controls the trade-off between user-individual and user-crowd multi-interest modeling. As synthesized users usually contain rich behaviors leading to high-computation of contrastive learning, we conduct backpropagation on $\mathcal{L}^{cst}$ every $\lfloor 1/\lambda \rfloor$ iterations ($\lambda > 0$) for efficiency consideration.

4 Model Complexity and Scalability

The computational complexity of LARMI comes from (a) LLM-driven inference and (b) multi-interest model training. For (a), assuming the complexity for analyzing each user's behaviors is $\mathcal{T}$,

Dataset	# Users	# Items	# Interactions	Avg Len	Density
Beauty	15,097	44,261	100,055	6.63	1.5e-4
Book	99,101	361,002	780,018	7.87	2.2e-5
Game	20,551	27,456	153,541	7.47	2.7e-4

Table 1: Dataset statistics.

the total cost for user-individual multi-interest extraction is $\mathcal{O}(M \cdot \mathcal{T})$. For user-crowd multi-interest extraction, it takes $\mathcal{O}(M^2)$ for MCP and $\mathcal{O}(Z \cdot \mathcal{T})$ for LLM-driven analysis, where $Z << M$. Therefore, the overall complexity for LLM-driven multi-interest analysis is $\mathcal{O}(M^2 + M \cdot \mathcal{T})$, which is similar to existing LLM-based recommendation methods (Dai et al., 2023; Hou et al., 2024). Regarding (b), computing collaborative multi-interests takes $\mathcal{O}(L \cdot K \cdot d)$, where L, K, d are sequence length, number of routing facets, embedding dimension. Aligning semantic and collaborative multi-interests also takes $\mathcal{O}(L \cdot K \cdot d)$. Therefore, the overall complexity of user-individual multi-interest learning is $\mathcal{O}(M \cdot L \cdot K \cdot d)$, which is equivalent to existing multi-interest recommendation models (Li et al., 2019; Xie et al., 2023). For contrastive learning, it takes $\mathcal{O}(Z \cdot \overline{|c_{(u')}|}^2 \cdot d)$ for each update, where $\overline{|c_{(u')}|}$ is the average number of behaviors for synthesized users. To tackle the high computational complexity in contrastive learning, we only update the loss every $\lfloor 1/\lambda \rfloor$ iteration. In summary, our method matches existing complexity and avoids online latency, thus ensuring scalability.

5 Experiments

5.1 Experimental Setup

Datasets: We use three subcategories of Amazon Review Data (Ni et al., 2019) with varying scales: *Beauty*, *Books*, and *Video Games*, abbreviated as Beauty, Book, and Game, respectively. All three datasets contain users' ratings on items with timestamps and item title information. Following prior studies (Xie et al., 2023; Du et al., 2024b), we filter out users and items with less than 5 records, then we convert ratings as implicit feedback for these three datasets. The statistical details of datasets are summarized in Table 1.

Evaluation Settings: To ensure a fair comparison, we follow prior studies (Xie et al., 2023), we chronologically split the user interactions with maximum length of 20 into training, validation, and test sets by the proportion of 6:2:2 and test last 20% items in each sequence. We adopt three widely used top-n evaluation metrics, i.e., Recall (R), Hit Rate (H), and Normalized Discounted Cumulative Gain (N), to evaluate all methods with $n = \{20, 50\}$.

Baseline Methods: We compare our model LARMI with the following baseline methods. **Pop** takes the most popular items as the recommendation results. **GRU4Rec** (Hidasi et al., 2016) models sequential behaviors through RNN structure. **LLM-BRec** (Harte et al., 2023) leverages an LLM to produce expressive embeddings by the BERT4Rec structure. **MIND** (Li et al., 2019) uses dynamic routing with a capsule network for multi-interest learning. **ComiRec-SA** (Cen et al., 2020) allows diversity control and introduces multi-head attention to model users' multi-interests. **Re4** (Zhang et al., 2022) leverages the backward flow to re-examine and regulate interest representations. **REMI** (Xie et al., 2023) introduces interest-aware hard negative sampling with routing variation regularization for multi-interest learning. **DisMIR** (Du et al., 2024b) formulates an item partition problem to encourage items in each group to focus on a discriminated interest. **EIMF** (Qiao et al., 2024) uses an LLM to extract similar items for multi-interest modeling.

5.2 Implementation Details

For a fair comparison, all methods are optimized by the Adam optimizer with a batch size of 128 and we adopt fixed embedding dimension 64 for all methods following (Xie et al., 2023; Du et al., 2024b). For all methods, we select the best performance by varying the number of interests in $\{2, 4, 6, 8\}$, learning rate in $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$, and weight decay in $\{1e^{-4}, 1e^{-5}, 1e^{-6}\}$. For hyper-parameters, we set $\tau = 0.1$ and $\lambda = 0.01$ for the contrastive loss $\mathcal{L}^{cst}$. For the MCP solver, we follow the same hyper-parameters as suggested in the original paper (Wang et al., 2022). For the LLM implementation, we use the *gpt-4o* as the backbone model with temperature set to 0 for reproducibility. For other hyper-parameters in baseline methods, we follow the authors' implementation if they exist, otherwise we tune them to their best according to their performance on the validation set.

5.3 Comparison with Baselines

From the results in Table 2, we summarize our key findings to answer RQ1. First, LARMI consistently outperforms baselines across all three datasets, highlighting the effectiveness of our LLM-driven analysis. In addition, the improvements demon-

	Metrics	Pop	GRU4Rec	LLMBRec	MIND	ComiRec	Re4	REMI	DisMIR	EIMF	**LARMI**
Beauty	*R@20*	0.0228	0.0349	0.0289	0.0477	0.0367	0.0550	0.0616	0.0702	0.0765	**0.0872**
	N@20	0.0161	0.0180	0.0205	0.0248	0.0176	0.0271	0.0320	0.0364	0.0390	**0.0443**
	H@20	0.0351	0.0454	0.0580	0.0669	0.0500	0.0715	0.0838	0.1057	0.1126	**0.1380**
	R@50	0.0391	0.0452	0.0473	0.0646	0.0519	0.0751	0.0817	0.0955	0.0982	**0.1092**
	N@50	0.0209	0.0186	0.0272	0.0257	0.0194	0.0274	0.0325	0.0433	0.0427	**0.0505**
	H@50	0.0593	0.0613	0.0943	0.0897	0.0702	0.0987	0.1099	0.1360	0.1429	**0.1552**
Book	*R@20*	0.0075	0.0215	0.0187	0.0236	0.0275	0.0298	0.0441	0.0639	0.0722	**0.0804**
	N@20	0.0052	0.0112	0.0132	0.0154	0.0166	0.0187	0.0293	0.0401	0.0413	**0.0485**
	H@20	0.0121	0.0314	0.0311	0.0334	0.0382	0.0420	0.0639	0.1034	0.1060	**0.1228**
	R@50	0.0133	0.0296	0.0256	0.0313	0.0381	0.0425	0.0592	0.0898	0.0951	**0.1036**
	N@50	0.0070	0.0113	0.0148	0.0158	0.0169	0.0194	0.0305	0.0404	0.0443	**0.0519**
	H@50	0.0217	0.0431	0.0538	0.0482	0.0570	0.0630	0.0915	0.1367	0.1436	**0.1594**
Game	*R@20*	0.0226	0.0751	0.0661	0.0950	0.0751	0.0967	0.1082	0.1221	0.1172	**0.1305**
	N@20	0.0139	0.0393	0.0420	0.0514	0.0384	0.0533	0.0543	0.0618	0.0604	**0.0713**
	H@20	0.0372	0.1099	0.1137	0.1401	0.1050	0.1465	0.1571	0.1689	0.1593	**0.2133**
	R@50	0.0435	0.1073	0.0849	0.1387	0.1145	0.1409	0.1510	0.1597	0.1571	**0.1742**
	N@50	0.0206	0.0423	0.0445	0.0552	0.0401	0.0558	0.0581	0.0689	0.0627	**0.0774**
	H@50	0.0710	0.1552	0.1576	0.2048	0.1496	0.2007	0.2175	0.2479	0.2315	**0.2662**

Table 2: Performance comparison of baseline methods and our proposed LARMI on three datasets. The best results are in **bold** and the runner-up results are underlined. The improvements are significant on the t-test ($p \leq 0.05$).

strate the advantage of integrating LLM analysis with our adaptive and representative multi-interest modeling framework by capturing diverse and meaningful user interests. We mitigate issues of uninterpretable agnostic granularity in LLM generation and the lack of a representative global perspective. Second, we observe that multi-interest baselines generally outperform single-interest baselines, showing that capturing multiple facets of user interests is beneficial for better results. Third, the superior performance of LARMI over the LLM-based baseline demonstrates that our approach achieves effective and coherent integration of the LLM-driven analysis and conventional multi-interest recommendation methods. Specifically, LARMI outperforms the relatively strong EIMF due to its emphasis on personalization and the granularity issue. Last, the relatively strong performance of DisMIR shows the positive impact of a global perspective item partition task. However, relying solely on the sparse co-occurrence of items limits the insights about item relationships. In comparison, LARMI produces synthesized users with compact and representative item subsets by formulating and solving the MCP, thus achieving improvements.

5.4 Ablation Studies

To validate the effects of key components, we conduct ablation studies as follows. **w/o-sem** removes the semantic-based multi-interest modeling with LLM-driven analysis at user-individual level, i.e., $\boldsymbol{h}_f^i = 0$. **w/o-col** removes the collaborative-based multi-interest modeling at the user-individual level, i.e., $\boldsymbol{o}_k^i = \boldsymbol{z}_k^i$. **w/o-com** removes the compactness rule for user synthesis, e.g., MCP formulation and user synthesis based on users' similarities. **w/o-rep** removes the representativeness rule for user synthesis. Instead, it randomly selects cliques as synthesized users for user-crowd level analysis.

	Metrics	w/o-sem	w/o-col	w/o-com	w/o-rep	LARMI
Beauty	*R@20*	0.0544	0.0604	0.0841	0.0850	**0.0872**
	N@20	0.0272	0.0310	0.0403	0.0425	**0.0443**
	H@20	0.0745	0.0920	0.1310	0.1344	**0.1380**
	R@50	0.0747	0.0889	0.1042	0.1058	**0.1092**
	N@50	0.0285	0.0338	0.0471	0.0479	**0.0505**
	H@50	0.1013	0.1279	0.1473	0.1525	**0.1552**
Book	*R@20*	0.0504	0.0585	0.0745	0.0783	**0.0804**
	N@20	0.0319	0.0362	0.0442	0.0461	**0.0485**
	H@20	0.0775	0.0891	0.1153	0.1197	**0.1228**
	R@50	0.0624	0.0740	0.0919	0.0937	**0.1036**
	N@50	0.0352	0.0416	0.0498	0.0510	**0.0519**
	H@50	0.1079	0.1263	0.1507	0.1558	**0.1594**
Game	*R@20*	0.0905	0.0934	0.1186	0.1256	**0.1305**
	N@20	0.0430	0.0468	0.0677	0.0695	**0.0713**
	H@20	0.1460	0.1523	0.2012	0.2084	**0.2133**
	R@50	0.1331	0.1405	0.1679	0.1708	**0.1742**
	N@50	0.0481	0.0525	0.0703	0.0750	**0.0774**
	H@50	0.1767	0.1953	0.2560	0.2589	**0.2662**

Table 3: Ablation study results on Beauty, Book and Game datasets. The best scores are in **bold**.

From Table 3, first, at the user-individual level, the lack of semantic multi-interests (w/o-sem) results in the worst performance, proving the effectiveness of LLM-driven analysis in handling the limitations of existing multi-interest modeling assumptions like co-occurring items simply indicating same interests. Second, w/o-col also shows inferior performance, showing our integration with collaborative interests and the alignment module

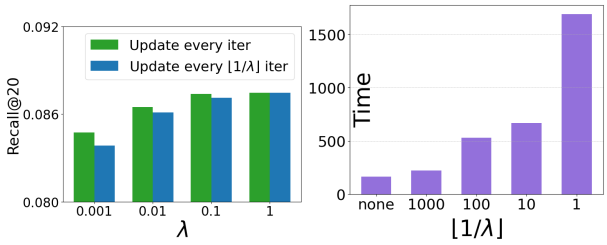

Figure 4: The model performance (a) and training time (b) with varying loss weights.

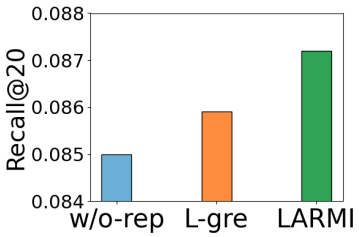

Figure 5: Model performance across MCP solve types.

is successful in alleviating the agnostic granularity issue in LLM-driven multi-interest analysis. Third, w/o-com and w/o-rep shows degraded performance, indicating that analyzing multi-interests for individual users only provides limited insights. On the one hand, generating user cliques from similar preferences ensures a moderate number of interests for synthesized users, thus each containing a rich set of cohesive items. On the other hand, the formulated MCP is crucial for selecting a representative subset that reduces redundancy and enhances representativeness. Thus, LARMI achieves representative multi-interest modeling by our user-crowd level multi-interest extraction.

5.5 Hyper-parameter Analysis

5.5.1 Loss Weight and Update Strategy

Figure 4 investigates the impact of (a) the loss weight λ on model accuracy and (b) the training time, with an update conducted every $\lfloor 1/\lambda \rfloor$ iterations. Higher λ (lower $\lfloor 1/\lambda \rfloor$) usually leads to higher model accuracy but requires a significantly longer time for model training. To balance effectiveness and efficiency, we suggest selecting $\lfloor 1/\lambda \rfloor = 100$. Generally, we observe that their performance improves as λ increases, with only slight differences between the two update strategies when $\lambda \geq 1 \times 10^{-2}$. Therefore, the efficient update strategy with a proper setting of λ significantly accelerates the training while maintaining comparable performance.

	Metrics	2	4	6	8
Beauty	*R@20*	0.0869	0.0872	0.0847	0.0849
	N@20	0.0438	0.0443	0.0423	0.0417
	H@20	0.1374	0.1380	0.1295	0.1334
	R@50	0.1077	0.1092	0.1009	0.1006
	N@50	0.0497	0.0505	0.0469	0.0486
	H@50	0.1530	0.1552	0.1523	0.1537
Book	*R@20*	0.0775	0.0804	0.0799	0.0783
	N@20	0.0479	0.0485	0.0497	0.0482
	H@20	0.1195	0.1228	0.1206	0.1190
	R@50	0.1004	0.1036	0.1042	0.0997
	N@50	0.0509	0.0519	0.0517	0.0483
	H@50	0.1588	0.1594	0.1598	0.1586
Game	*R@20*	0.1251	0.1305	0.1274	0.1240
	N@20	0.0704	0.0713	0.0717	0.0711
	H@20	0.2119	0.2133	0.2115	0.2084
	R@50	0.1689	0.1742	0.1754	0.1712
	N@50	0.0732	0.0774	0.0741	0.0748
	H@50	0.2601	0.2662	0.2620	0.2635

Table 4: LARMI performance of different interest nums.

5.5.2 MCP Solver

In Figure 5, we test the performance of different solvers in MCP for the multi-interest extraction at the user-crowd level. Specifically, we replace the MCP solver (Wang et al., 2022) with a greedy search of representative synthesized users, L-gre. First, LARMI shows an advantage over L-gre, reflecting the effectiveness of the MCP solver in selecting representative synthesized users. Second, both of the two approaches (LARMI and L-gre) outperform the w/o-rep variant, indicating the necessity of selecting representative synthetic users at the user-crowd level.

5.5.3 Number of Interests

Table 4 shows the effect of the number of multi-interests K for LARMI, where the optimal number of interests is 4 for all three datasets, matching the average numbers of LLM-derived clusters. Specifically, insufficient interest numbers make it hard to capture the diverse facets of user preferences, while too large interest numbers may lead to the modeling outcomes with overly-fine clusters. As a result, we suggest a moderate interest number $K = 4$ for real-world applications.

5.6 Adaptive Granularity Control Analysis

To empirically validate how our model adaptively addresses the agnostic granularity issue with raw LLM outputs, we visualize the alignment module's behavior. Specifically, we define cluster coarseness (x-axis) as the semantic divergence of a raw LLM-derived semantic cluster $\mathcal{C}_f^i$, computed by the average pairwise distance of the item embeddings

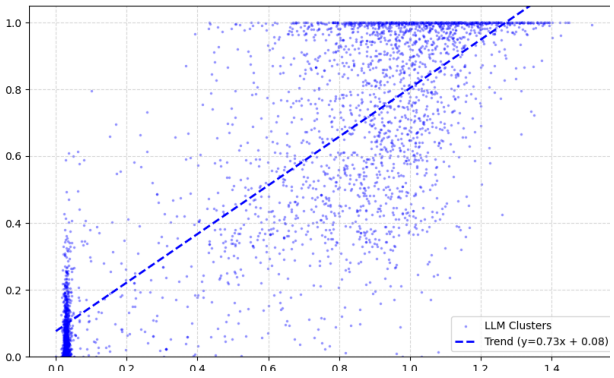

Figure 6: Positive correlation occurs between LLM cluster coarseness (x-axis) and the alignment module's attention concentration (y-axis).

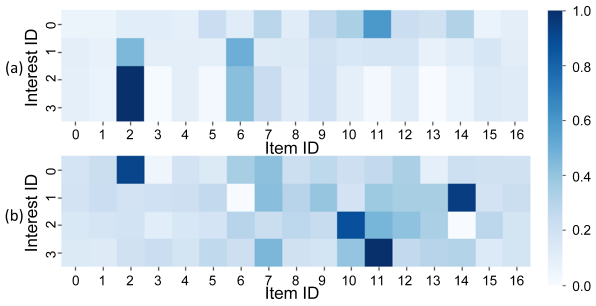

Figure 7: Multi-interest heatmap of baseline method MIND (a) and our proposed LARMI (b).

within that cluster. Interest Concentration (y-axis) measures the sharpening effect of our alignment module, which is computed according to the negative entropy of the attention distribution a_j.

As shown in Figure 6, there is a strong positive correlation between the two metrics, demonstrating our module's adaptive capability. Coarse clusters exhibit high attention concentration, indicating that the model selectively emphasizes specific items within a broad semantic group to filter noise and align with the user's certain interest. Conversely, over-fine clusters are effectively merged into a unified, broader representation for further alignment with collaborative signals. Thus, LARMI successfully mitigates the issue of agnostic LLM semantics, ensuring multi-interest profiles are dynamically adjusted to the optimal level of granularity.

5.7 Case Study

We illustrate a case study to qualitatively investigate the effect of our model for multi-interest extraction via LLMs. To visualize the cosine similarity between interests and items, we plot the multi-interest heatmap in Figure 7 for a classic baseline MIND and our LARMI model, which corresponds to a user sequence randomly sampled from *Beauty* dataset containing 17 items across 4 interests. In the heatmap, darker colors indicate higher cosine similarity between interests and items. We notice that the heatmaps with interest ID 2 and 3 in MIND exhibit significant overlap, indicating collapsed facets. Rather than being dominated mainly by one item for each interest, each interest in LARMI contains multiple relevant items, indicating discriminated and balanced multi-interest modeling. This improvement validates LARMI's ability to adaptively capture diverse and fine-grained interests with the LLM and our model design.

6 Conclusion

In this paper, we present LARMI, the LLM-based Adaptive and Representative Multi-Interest approach that leverages the semantic knowledge and reasoning skills of LLMs to address critical challenges in multi-interest modeling for recommendation. At the user-individual level, LARMI provides an adaptive solution to the agnostic granularity in raw LLM generations, which merges and aligns semantic clusters with collaborative interests. At the user-crowd level, LARMI leverages MCP optimization and contrastive learning to mitigate the limitation in individual user behaviors and extend the scope of LLM analysis to a representative global perspective. Extensive experiments validate the superiority of LARMI over single- and multi-interest baselines. Further analysis supports the effectiveness of our model design. Future work includes fine-tuning open-source LLMs to align semantics with collaborative behaviors to further improve the scalability for multi-interest modeling.

Acknowledgements

This research is supported in part by the Singapore MOE AcRF Tier 1 funding (RG16/25); and in part by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 1 grant, and funded through the SMU-SUTD Internal Research Grant Call (SMU-SUTD 2023_02_01); and in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP201230015).

Limitations

The limitations in our current framework include the following points. First, our method relies on the

semantic richness of item metadata to prompt the LLM for multi-interest analysis. However, it may be less effective if the data has noise or incompleteness, whereas ID-based methods would be unaffected. Second, our reliance on external APIs may be unstable due to changes in backbones. Third, the inference cost of querying Large Language Models remains higher than traditional lightweight recommendation models, especially with larger datasets.

References

Keqin Bao, Ming Yan, Yang Zhang, Jizhi Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2025. Customizing in-context learning for dynamic interest adaption in LLM-based recommendation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14278–14291.

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. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, page 1007–1014.

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, page 2942–2951.

Zheng Chai, Zhihong Chen, Chenliang Li, Rong Xiao, Houyi Li, Jiawei Wu, Jingxu Chen, and Haihong Tang. 2022. User-aware multi-interest learning for candidate matching in recommenders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, page 1326–1335.

Gaode Chen, Xinghua Zhang, Yanyan Zhao, Cong Xue, and Ji Xiang. 2021. Exploring periodicity and interactivity in multi-interest framework for sequential recommendation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1426–1433.

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. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, page 1126–1132.

Yingpeng Du, Di Luo, Rui Yan, Xiaopei Wang, Hongzhi Liu, Hengshu Zhu, Yang Song, and Jie Zhang. 2024a. Enhancing job recommendation through llm-based generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8363–8371.

Yingpeng Du, Ziyan Wang, Zhu Sun, Yining Ma, Hongzhi Liu, and Jie Zhang. 2024b. Disentangled multi-interest representation learning for sequential recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, page 677–688.

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 *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, page 299–315.

Qing Guo, Zhu Sun, Jie Zhang, and Yin-Leng Theng. 2020. An attentional recurrent neural network for personalized next location recommendation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 34, pages 83–90.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, page 1025–1035.

Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval (ECIR)*, pages 364–381.

Yangqin Jiang, Yuhao Yang, Lianghao Xia, Da Luo, Kangyi Lin, and Chao Huang. 2025. RecLM: Recommendation instruction tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15443–15459.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206.

Samir Khuller, Anna Moss, and Joseph Seffi Naor. 1999. The budgeted maximum coverage problem. *Information processing letters*, 70(1):39–45.

Yankun Le, Haoran Li, Baoyuan Ou, Yingjie Qin, Zhixuan Yang, Ruilong Su, and Fu Zhang. 2025. Diffusion model for interest refinement in multi-interest recommendation. *Preprint*, arXiv:2502.05561.

Jaeri Lee, Jeongin Yun, and U Kang. 2024. Towards true multi-interest recommendation: Enhanced scheme for balanced interest training. In *2024 IEEE International Conference on Big Data (BigData)*, pages 394–402.

Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, page 2615–2623.

Danyang Liu, Yuji Yang, Mengdi Zhang, Wei Wu, Xing Xie, and Guangzhong Sun. 2022. Knowledge enhanced multi-interest network for the generation of recommendation candidates. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, page 3322–3331.

Yaokun Liu, Xiaowang Zhang, Minghui Zou, and Zhiyong Feng. 2024. Attribute simulation for item embedding enhancement in multi-interest recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, page 482–491.

Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, page 483–491.

Fnu Mohbat and Mohammed J Zaki. 2025. KERL: Knowledge-enhanced personalized recipe recommendation using large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19125–19141.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Shutong Qiao, Chen Gao, Yong Li, and Hongzhi Yin. 2024. Llm-assisted explicit and implicit multi-interest learning framework for sequential recommendation. *arXiv preprint arXiv:2411.09410*.

Zekai Qu, Ruobing Xie, Chaojun Xiao, Zhanhui Kang, and Xingwu Sun. 2024. The elephant in the room: Rethinking the usage of pre-trained language model in sequential recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys)*, page 53–62.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Rlmrec: Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024 (TheWebConf)*, pages 3464–3475.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, page 3859–3869.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, page 890–896.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, page 1441–1450.

Yu Tian, Jianxin Chang, Yanan Niu, Yang Song, and Chenliang Li. 2022. When multi-level meets multi-interest: A multi-grained neural model for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, page 1632–1641.

Runzhong Wang, Li Shen, Yiting Chen, Xiaokang Yang, Dacheng Tao, and Junchi Yan. 2022. Towards one-shot neural combinatorial solvers: Theoretical and empirical notes on the cardinality-constrained case. In *The 11th International Conference on Learning Representations (ICLR)*.

Xinfeng Wang, Jin Cui, Fumiyo Fukumoto, and Yoshimi Suzuki. 2025. AGRec: Adapting autoregressive decoders with graph reasoning for LLM-based sequential recommendation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7076–7090.

Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Zhang, Qing Cui, and 1 others. 2024. Llmrg: Improving recommendations through large language model reasoning graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 19189–19196.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 806–815.

Haotian Wu, Yingpeng Du, Tianjun Wei, Puay Siew Tan, Jie Zhang, Ong Yew Soon, and Zhu Sun. 2026. Efficient large language models for recommendation: A survey. *TechRxiv*, 2026(0126).

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

Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Xiyue Zhang, Hongsheng Yang, Jian Pei, and Liefeng Bo. 2021. Knowledge-enhanced hierarchical graph transformer network for multi-behavior recommendation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, pages 4486–4493.

Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020. Deep multi-interest network for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, page 2265–2268.

Yueqi Xie, Jingqi Gao, Peilin Zhou, Qichen Ye, Yining Hua, Jae Boum Kim, Fangzhao Wu, and Sunghun Kim. 2023. Rethinking multi-interest learning for candidate matching in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, page 283–293.

Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2022. Re4: Learning to re-contrast, re-attend, reconstruct for multi-interest recommendation. In *The ACM Web Conference (TheWebConf)*, pages 2216–2226.

Shuai Zhang, Yi Tay, Lina Yao, Aixin Sun, and Jake An. 2019. Next item recommendation with self-attentive metric learning. In *Thirty-Third AAAI conference on artificial intelligence (AAAI)*, volume 9.

Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. 2023. Generative job recommendations with large language model. *arXiv preprint arXiv:2307.02157*.

A Backgrounds

A.1 Multi-Interest Modeling for Recommendation

Capturing different preference representations of users is essential for multi-interest methods. Capsule networks (Sabour et al., 2017) have gained popularity for multi-interest modeling recently (Li et al., 2019; Xie et al., 2023). Specifically, the capsule network $CapsuleNet(\cdot)$ can generate users' K-interest representations $\{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_K\}$ derived from their sequential behaviors $s_{(u_i)}$, where K denotes the pre-defined number of users' multi-interests. The k-th collaborative interest capsule $\boldsymbol{m}_k \in \mathbb{R}^d$ is calculated by:

$$\boldsymbol{m}_k = \sum_{j=1}^{L} b_{jk} \cdot \boldsymbol{W} \cdot \boldsymbol{v}_j^i, \quad k = 1, \cdots, K, \tag{16}$$

where $\boldsymbol{v}_j^i \in \mathbb{R}^d$ denotes the embeddings of the j-th item in the user u_i's behaviors $s_{(u_i)}$, and d denotes the dimension of embedding space. $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is the transformation matrix, and b_{jk} denotes the routing weight of item v_j^i to the k-th interest capsule. The routing weight b_{jk} is calculated by the softmax operation of the routing logits g_{jk} that measures the similarity between the item embedding and squashed vector $\boldsymbol{e}_k$, i.e.,

$$\begin{aligned} \boldsymbol{e}_k &= \frac{||\boldsymbol{m}_k||^2}{||\boldsymbol{m}_k||^2 + 1} \cdot \frac{\boldsymbol{m}_k}{||\boldsymbol{m}_k||}, \\ g_{jk} &\leftarrow g_{jk} + \boldsymbol{v}_j^{i\top} \cdot \boldsymbol{W} \cdot \boldsymbol{e}_k, \\ b_{jk} &= \frac{exp(g_{jk})}{\sum_{k=1}^{K} exp(g_{jk})}, \end{aligned} \tag{17}$$

where the iterative process updates the routing logits g_{jk} and recalculates the routing weights b_{jk} based on updated $\boldsymbol{m}_k$.

A.2 Max Covering Problem

The max covering problem (MCP) (Khuller et al., 1999) is a combinatorial optimization problem, widely applied in decision-making under uncertainty. Given P sets, each set containing an indefinite number of objects, and Q objects, each object associated with a specific value, the MCP aims to select a subset of Z sets ($Z \ll P$) such that the union of Z sets maximizes the sum of the associated values of the covered objects. The problem can be formulated as:

$$\begin{aligned} &\max_{\boldsymbol{x}} \sum_{j=1}^{Q} \left(\mathbb{I}\left(\sum_{i=1}^{P} \boldsymbol{x}_i \boldsymbol{A}_{ij} \geq 1 \right) \cdot w_j \right), \\ &\text{s.t.} \quad \boldsymbol{x} \in \{0,1\}^P, \ \|\boldsymbol{x}\|_0 \leq Z, \end{aligned} \tag{18}$$

where $\boldsymbol{A} \in \{0,1\}^{P \times Q}$ is the adjacency matrix of a bipartite graph linking the sets and objects, $w_j \in \mathbb{R}$ denotes the j-th object value, $\mathbb{I}(\cdot)$ is a condition indicator function, $\boldsymbol{x}$ is the selection outcome, and each element $\boldsymbol{x}_i$ is a scalar indicating whether the i-th set is selected in solution. To tackle the MCP, an advanced neural solver encodes the bipartite graph with a three-layer GraphSage model (Hamilton et al., 2017), integrates the MCP constraints into a differentiable layer, and then predicts the probabilities of selecting each set.