

Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation

Bowen Zheng
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
bwzheng0324@ruc.edu.cn

Yupeng Hou
University of California San Diego
La Jolla, United States
yphou@ucsd.edu

Hongyu Lu
WeChat, Tencent
Guangzhou, China
luhy94@gmail.com

Yu Chen
WeChat, Tencent
Beijing, China
nealcui@tencent.com

Wayne Xin Zhao✉
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
batmanfly@gmail.com

Ming Chen
WeChat, Tencent
Guangzhou, China
mingchen@tencent.com

Ji-Rong Wen†
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
jrwen@ruc.edu.cn

Abstract—Recently, large language models (LLMs) have shown great potential in recommender systems, either improving existing recommendation models or serving as the backbone. However, there exists a large semantic gap between LLMs and recommender systems, since items to be recommended are often indexed by discrete identifiers (*item ID*) out of the LLM’s vocabulary. In essence, LLMs capture language semantics while recommender systems imply collaborative semantics, making it difficult to sufficiently leverage the model capacity of LLMs for recommendation.

To address this challenge, in this paper, we propose a new LLM-based recommendation model called *LC-Rec*, which can better integrate language and collaborative semantics for recommender systems. Our approach can directly generate items from the entire item set for recommendation, without relying on candidate items. Specifically, we make two major contributions in our approach. For item indexing, we design a learning-based vector quantization method with uniform semantic mapping, which can assign meaningful and non-conflicting IDs (called *item indices*) for items. For alignment tuning, we propose a series of specially designed tuning tasks to enhance the integration of collaborative semantics in LLMs. Our fine-tuning tasks enforce LLMs to deeply integrate language and collaborative semantics (characterized by the learned item indices), so as to achieve an effective adaptation to recommender systems. Extensive experiments demonstrate the effectiveness of our method, showing that our approach can outperform a number of competitive baselines including traditional recommenders and existing LLM-based recommenders. Our code is available at <https://github.com/RUCAIBox/LC-Rec/>.

Index Terms—Large Language Model, Semantic Integration, Sequential Recommendation

I. INTRODUCTION

Nowadays, recommender systems have become an essential part of various application platforms, aiming to recommend potential information resources to users based on their specific preferences. Since user preferences dynamically evolve over time, *sequential recommendation* has attracted great research

attention due to its advantages in capturing the sequential characteristics of user behaviors. To develop sequential recommenders, existing recommendation models [1], [2] are mostly built on sequential formatting of user interaction logs, taking *item ID* as the basic unit.

In existing literature, sequential recommenders adopt various deep neural networks to model user historical behaviors represented by item ID sequences, including RNN [1], [3], CNN [4], [5], GNN [6], [7], and Transformer [2], [8]. In addition to collaborative semantics within users’ historical behaviors, some studies [9]–[11] also try to enhance item sequence modeling by leveraging content information (*e.g.*, title, description, category). Moreover, pre-trained language models (PLMs) have also been employed for capturing the textual semantics reflected in item texts [12]–[15], to improve the recommendation performance.

Recently, the emergence of large language models (LLMs) has triggered a significant revolution in the research community. LLMs have shown great potential in various language based tasks, due to their excellent capabilities in semantic understanding and generation [16]. Specifically, there are also several attempts [17] that adapt LLMs for recommender systems (RS), to improve the item ranking performance [18] or boost the comprehensive recommendation capacities [19]. To develop capable LLM-based recommendation models, a fundamental challenge is that there exists a large gap between the *language semantics* modeled by LLMs and *collaborative semantics* implied by recommender systems. The key point is that, in existing recommendation models, user behaviors are often formatted into item ID sequences (possibly with feature IDs), but not textual descriptions. In other words, language models and recommendation models indeed employ two different vocabularies (token IDs *v.s.* item IDs) to learn their own semantic spaces. Such a semantic gap makes it difficult to sufficiently leverage the model capacity of LLMs for tackling the recommendation tasks.

✉ Corresponding author.

† Ji-Rong Wen is also with School of Information, Renmin University of China.

To address this issue, existing efforts can be divided into two main approaches. The first approach [18]–[21] verbalizes the user behaviors into text sequences (*e.g.*, concatenating the titles and category labels of the interacted items), and designs special prompts to instruct LLMs for fulfilling the recommendation tasks. Such an approach only captures limited item information (only considering language semantics), and can't guarantee the generation of in-domain items (relying on a candidate set). As another alternative approach, several studies [22], [23] design special item indexing mechanisms for building item vocabulary, and then learn to generate the target item for recommendation. However, given the large semantic gap, *simple* (*e.g.*, vocabulary building with vanilla item IDs) or *shadow* integration (*e.g.*, fine-tuning only with the target task) would be less effective to adapt LLMs for recommender systems.

Considering these issues, we aim to design a more effective semantic integration approach for developing LLM-based recommendation models. We tackle this semantic integration problem in two main aspects, namely item indexing and alignment tuning¹. For item indexing, an ideal allocation mechanism should produce *meaningful* (capturing item similarities), *unique* (without allocation conflicts), and *extensible* (generalizable to new items) IDs for effectively representing the items. For alignment tuning, it should be able to sufficiently integrate language semantics with collaborative semantics in LLMs, but not superficially fit the target recommendation task. Overall, our goal is to effectively establish the connections between the two kinds of different semantics and fully leverage the model capacity of LLMs for sequential recommendation.

To this end, in this paper, we propose **LC-Rec**, a new approach to integrate **L**anguage and **C**ollaborative semantics for improving LLMs in **R**ecommender systems. Our approach is built in a generative manner, where the recommendation task is cast into a token generation task as well. To achieve this, the key point lies in the semantic integration between language and collaborative semantics, so that LLMs can make the item recommendations just like they generate normal text contents. Our approach has made two major contributions in the aforementioned two aspects. For item indexing, we propose a tree-structured vector quantization (VQ) method to index the items with discrete IDs (called *item indices*). These item indices are learned based on the text embeddings of items encoded by LLMs, enabling the learned IDs to capture the intrinsic similarity among items. However, original VQ methods are likely to assign the same IDs to multiple items, which should be avoided in recommender systems. To tackle this problem, we further design a uniform semantic mapping method to mitigate the potential conflicts in ID allocation. For alignment tuning, we design a series of specific tasks to fine-tune LLMs for achieving semantic integration. In addition to the sequential item prediction, we consider both *explicit index-language alignment* and *implicit recommendation-oriented*

alignment. Our fine-tuning tasks enforce LLMs to deeply integrate language and collaborative semantics, so as to achieve an effective adaptation to recommender systems.

To evaluate our approach, we conduct extensive experiments on three real-world datasets. Our method achieves the best performance compared to a number of competitive baselines. Experimental results demonstrate that our approach can effectively align language and collaborative semantics via specially learned item indices, thereby significantly improving the recommendation performance. The contributions of this work can be summarized as follows:

- We present LC-Rec, a LLM-based sequential recommendation model, by effectively integrating language and collaborative semantics. LC-Rec can fulfill the sequential recommendation task in an autoregressive generation way, without relying on candidate sets.
- Our approach is built on a specially designed VQ method, which can capture item similarity and avoid ID conflicts in index allocation. Further, we propose a series of carefully designed tuning tasks for achieving effective semantic integration via item indices.
- We implement our method based on LLaMA [25] with 7B parameters. Extensive experiments on three public datasets demonstrate the effectiveness of our approach in integrating collaborative semantics into LLMs. The proposed method LC-Rec achieves an average performance improvement of 22% in full ranking evaluations, compared to all baseline methods.

II. RELATED WORK

A. Sequential Recommendation

Sequential recommendation aims to infer user preferences by analyzing historical interactions and predict the next item that would be suitable for that user [1], [2], [8], [26]. Many early methods are frequently based on Markov Chains techniques [27], [28]. Recently, typical methods become to adopt various deep neural networks to model user historical behaviors represented by item ID sequences, including RNN [1], [3], CNN [4], [5], GNN [6], [7], and Transformer [2], [8]. However, these methods only capture the collaborative relationship between items from user-item interactions, while ignoring the additional information rich in the item content information (*e.g.*, title, description, category). Therefore, several studies are devoted to utilizing additional information associated with items to enhance ID sequence modeling [9]–[11]. Furthermore, the inherent natural language characteristics of the item title and description have motivated numerous researchers to explore the utilization of pre-trained language models (PLMs) as feature encoders in recommender systems [12]–[15], [29]. In this paper, we aim to combine LLMs and recommendation tasks in a more effective way, leveraging LLMs as the recommender, which is reached through the proposed new item indexing and alignment tuning methods.

¹Note that the words “align” and “alignment” mainly refer to the integration between language semantics and collaborative semantics, but not what it means in human alignment [24] that instructs LLMs to follow human values or preferences.

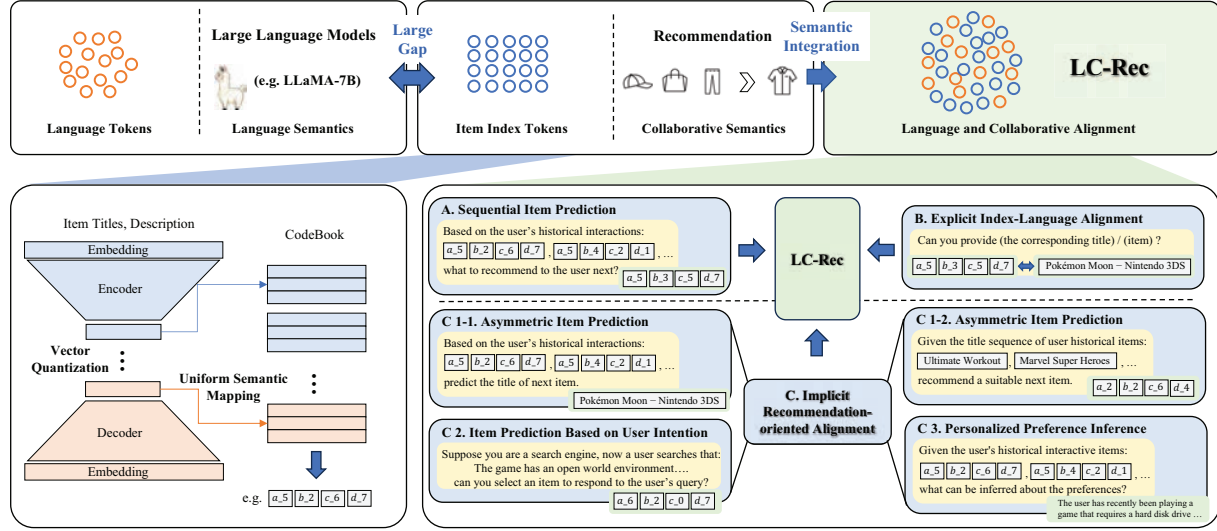


Fig. 1. The overall framework of our LC-Rec. We enhance language learning models (LLMs) by integrating language and collaborative semantics based on item indexing and alignment tuning, thereby adapting LLMs to recommender systems.

B. Large Language Models for Recommendation

Recently, large language models (LLMs) have gained significant popularity, with a wide range of applications spanning various domains of artificial intelligence [16], [30]–[33]. This is largely attributed to their superior capabilities in language semantic understanding and generation. In the context of RS, researchers have been working on adapting LLMs for RS to improve recommendation performance. A common approach is to represent user behaviors as text sequences (e.g., by concatenating the titles of historical items), and then design prompts to guide LLMs to perform the recommendation task [34]–[38].

However, a major challenge remains: there is a large gap between the language semantics modeled by LLMs and the collaborative semantics implied by recommender systems, which cannot be bridged by simple prompt design alone. To address this problem, existing efforts can be categorized into two main approaches. The first approach is to fine-tune the LLMs with text-based user behavior sequences [19]–[21], [39], [40]. However, these methods cannot guarantee the generation of in-domain items. Due to the constraints of limited context window size, these methods can only rank on a given candidate set, and can hardly be applied in a full ranking scenario. The second approach maintains the use of item IDs or introduces unique item indexing mechanisms. Given pure item ID sequences, models are trained to directly generate target item IDs for recommendation [22], [23], [41]. Although collaborative semantics between item indices are established, the language semantics modeled by LLMs and these item indices are not well aligned.

Additionally, we are aware of some concurrent works [42]–[44], which also recognize the issue of the large semantic gap between recommendation tasks and natural language tasks.

For instance, TransRec [42] employs multi-facet identifiers, combining ID, title, and attributes to balance item distinctiveness and semantics. CoLLM [43] incorporates collaborative semantics into LLMs by using representations of an external collaborative model as part of the input. CLLM4Rec [44] learns dual user/item embeddings based on recommendation task and content (e.g., reviews) generation task, respectively. A mutually-regularization loss is introduced for interaction between these two kinds of embeddings. Only the recommendation-task embeddings are for the final recommendation.

Contrasting these methodologies, our focus is to develop a deep and unified integration of language and collaborative semantics within LLMs through carefully crafted item indices. In particular, our approach uses a tree-structured vector quantization method to construct item indices. This method allows for better semantic integration by adding a small number (usually $\sim 1,000$) of additional tokens to LLMs. Furthermore, we introduce a series of semantic alignment tasks to fine-tune LLMs, aiming to achieve unified semantic integration in a practical recommendation setting (e.g., full ranking).

III. METHODOLOGY

In this section, we present the proposed LLM-based recommendation model **LC-Rec**, which integrates **Language** and **Collaborative** semantics for improving LLMs in **Recommender** systems.

A. Overview of the Approach

As we discussed in Section I, there exists a large gap between the *language semantics* modeled by LLMs and *collaborative semantics* implied by recommender systems, which limits the capacities of LLMs in recommender systems. To effectively bridge this gap, we consider enhancing the semantic integration in two major aspects.

- For *item indexing* (Section III-B), we represent an item with several learned discrete IDs via vector quantization based on text embeddings by LLMs, and further propose a uniform semantic mapping method to mitigate the potential conflicts in index assignment. In this way, the learned item indices can capture similarities between the textual semantics of item information, and provide a unique indexing representation for a specific item.
- For *alignment tuning* (Section III-C), we design a series of specific tuning tasks that enhance the integration between language semantics and collaborative semantics, not limited to the target recommendation task. Our approach can effectively integrate the collaborative semantics into LLMs, and sufficiently leverage the powerful model capacity of LLMs for recommendation tasks.

The overall framework of the proposed approach is shown in Figure 1. Next, we will present the details of our method.

B. Learning Item Indices for Semantic Integration

To extend the capacities of LLMs for recommendation, a fundamental problem is how to represent an item with index IDs (called *item indices*) and integrate these item indices into LLMs. We don't adopt the original item ID (resulting in a very large vocabulary), but instead employ vector quantization techniques to represent an item with a small number of discrete indices. These indices are constructed by leveraging relevant item information (e.g., item text representations), and the token embeddings associated with these discrete indices can be further optimized to fit the recommendation task (Section III-C). In this part, we present the approach for learning item indices for subsequent semantic integration. The approach consists of two major steps: it first conducts vector quantization based on text embeddings of items, so that the original representations of item indices can capture latent textual semantic correlations between items; then, it proposes a uniform semantic mapping to mitigate the potential conflicts in item index assignment. Next, we introduce the two parts in detail.

1) *Vector Quantization Indexing*: In recommender systems, it is common to associate each item with a single unique ID (called vanilla ID). However, it would directly introduce a large vocabulary of item IDs when dealing with a great number of items (i.e., a large item set). Further, such an approach is easy to suffer from the OOV issue when adapting to new items (e.g., cold-start items). To address this issue, we borrow the idea of existing studies [15], [23], [45] to learn indices associated with latent semantics for items. Specifically, each item is represented by a composition of discrete indices corresponding to its own latent semantic, and each discrete index can be shared by multiple items. The basic idea is that similar items tend to be assigned with a portion of common semantic indices, such that each unique semantic index can be aligned to some kind of latent semantics.

To derive these semantic indices, we first employ LLMs (e.g., LLaMA) to encode the attached text information for an item, and obtain the text embeddings as the initial item representation. Further, we propose to use a vector quantization approach to

create discrete indices based on item embeddings. Specifically, we take the item embeddings encoded by LLMs as input, and then train a Residual-Quantized Variational AutoEncoder (RQ-VAE) for generating item indices. RQ-VAE [46] is a multi-level vector quantizer, which recursively quantized the residual vectors from coarse to fine to generate a set of codewords (i.e., item indices). For an item embedding e , RQ-VAE first encodes it into a latent representation z . At each level h , we have a codebook $\mathcal{C}^h = \{v_k^h\}_{k=1}^K$, where each codebook vector v_k^h is a learnable cluster center. Then the residual quantization process can be expressed as:

$$c_i = \arg \min_k \|r_i - v_k^i\|_2^2, \quad (1)$$

$$r_{i+1} = r_i - v_{c_i}^i, \quad (2)$$

where c_i is the i -th codeword of the item indices and r_i is the residual vector in the i -th RQ level, and we set $r_1 = z$.

When we have H -level codebooks, the quantization representation of z can be obtained according to $\hat{z} = \sum_{i=1}^H v_{c_i}^i$. Then \hat{z} will be used as decoder input to reconstruct the item embedding e . The overall loss function is as follows:

$$\mathcal{L}_{\text{RECON}} = \|e - \hat{e}\|_2^2, \quad (3)$$

$$\mathcal{L}_{\text{RQ}} = \sum_{i=1}^H \|\text{sg}[r_i] - v_{c_i}^i\|_2^2 + \beta \|r_i - \text{sg}[v_{c_i}^i]\|_2^2, \quad (4)$$

$$\mathcal{L}_{\text{RQ-VAE}} = \mathcal{L}_{\text{RECON}} + \mathcal{L}_{\text{RQ}}, \quad (5)$$

where \hat{e} is the output of the decoder, $\text{sg}[\cdot]$ represents the stop-gradient operator, and β is a loss coefficient, usually set to 0.25. The overall loss is divided into two parts, $\mathcal{L}_{\text{RECON}}$ is the reconstruction loss, and \mathcal{L}_{RQ} is the RQ loss used to minimize the distance between codebook vectors and residual vectors.

Compared with traditional VQ approaches, RQ offers the advantage of achieving a larger expression space with a smaller codebook size [46], [47]. Besides, its coarse-to-fine quantification method results in a tree-structured item index, which is beneficial for autoregressive generation. In fact, the RQ approach has demonstrated its effectiveness across various autoregressive generation tasks, such as autoregressive image generation [47] and generative recommendation [45]. Instead of simply employing VQ for item indexing [15], [45], we consider two key improvements for deriving meaningful item indices. First, there should be no conflicts in item indices, which is a common issue with VQ but should not occur in recommender systems. Second, the established semantic spaces of item indices should be aligned with the semantics of LLMs, in order to better leverage the powerful model capacity of LLMs for recommendation. We next introduce the two major improvements in our approach.

2) *Conflict Mitigation via Uniform Semantic Mapping*: Since we adopt the tree structure for learning item indices, it might lead to index conflicts among items within the same leaf node. To address this issue, existing solutions [23], [45] typically add an additional layer to the index tree and assign a distinct supplementary index ID to each item in a node with conflicts. However, this approach introduces semantically

Algorithm 1 RQ with Uniform Semantic Mapping

Input: Batch item representations $\mathbf{B} = \{\mathbf{z}^n\}_{n=1}^{|B|}$; H -level codebooks $\{\mathbf{C}^h\}_{h=1}^H$.

Output: Item indices $\{[c_1^n, c_2^n, \dots, c_H^n]\}_{n=1}^{|B|}$; Quantified representations $\{\hat{\mathbf{z}}^n\}_{n=1}^{|B|}$.

- 1: Let initial residual vectors $\mathbf{r}_1^n = \mathbf{z}^n, \forall \mathbf{z}^n \in \mathbf{B}$
- 2: **for** $i = 1$ **to** H **do**
- 3: **if** $i < H$ **then**
- 4: Solve $\{c_i^n\}_{n=1}^{|B|}$ according to Eqn. (1)
- 5: **else**
- 6: Solve $\{c_H^n\}_{n=1}^{|B|}$ according to Eqn. (6) via Sinkhorn-Knopp algorithm
- 7: **end if**
- 8: Obtain $\{\mathbf{r}_{i+1}^n\}_{n=1}^{|B|}$ according to Eqn. (2)
- 9: **end for**
- 10: **for all** $\mathbf{z}^n \in \mathbf{B}$ **do**
- 11: Calculate quantified representations by $\hat{\mathbf{z}}^n = \sum_{i=1}^H \mathbf{v}_{c_i^n}^i$
- 12: **end for**
- 13: **return** $\{[c_1^n, c_2^n, \dots, c_H^n]\}_{n=1}^{|B|}$ and $\{\hat{\mathbf{z}}^n\}_{n=1}^{|B|}$

irrelevant distributions in the tree's final layer. Additionally, these newly integrated IDs might also affect the original item representations.

Considering these issues, we propose a new conflict mitigation method to avoid the clustering of multiple items within the same leaf node. Our objective is to ensure that item semantics are uniformly distributed across different codebook embeddings at the last index level. To achieve this, we introduce a uniform distribution constraint to the original formulation:

$$\begin{aligned}
& \min \sum_{\mathbf{r}_H \in \mathbf{B}} \sum_{k=1}^K q(c_H = k | \mathbf{r}_H) \|\mathbf{r}_H - \mathbf{v}_k^H\|_2^2, \\
& \text{subject to: } \sum_{k=1}^K q(c_H = k | \mathbf{r}_H) = 1, \\
& \sum_{\mathbf{r}_H \in \mathbf{B}} q(c_H = k | \mathbf{r}_H) = \frac{|B|}{K},
\end{aligned} \tag{6}$$

where \mathbf{B} is a batch of residual vectors in the last index level. Following [48]–[51], by considering $\|\mathbf{r}_H - \mathbf{v}_k^H\|_2^2$ as the cost of semantic mapping, this problem can be viewed as an optimal transmission problem. In this setting, $q(c_H = k | \mathbf{r}_H)$ represents the transmission or mapping scheme that needs to be solved. In our implementation, we solve this equation by Sinkhorn-Knopp algorithm [52]. The overall process of RQ with uniform semantic mapping is shown in Algorithm 1.

By optimizing the loss in Eqn. (5), we can obtain a trained encoder and multi-level codebooks. During the construction of item indices, we first generate indices based on Eqn. (1). After that, for each group of conflicting items, the codewords of these items at the last level will be redistributed uniformly based on Eqn. (6). Such a two-stage process can also improve efficiency and reduce unnecessary random noise introduced by batching items [48].

C. Aligning Language and Collaborative Semantics in LLMs

After learning item indices, a straightforward approach is to integrate these index IDs into the LLM vocabulary, so that LLM can fulfill the recommendation task in a generative way that gradually omits the indices of items. However, these item indices are essentially OOV tokens for LLMs, and it is necessary to conduct the alignment between language and collaborative semantics. For this purpose, we design a series of semantic alignment tasks to assign language and collaborative semantics for tuning LLMs, including the primary objective of sequential item prediction, explicit index-language alignment (identifying the corresponding item via their indices), and implicit recommendation-oriented alignment (enhancing comprehension of the language and collaborative semantics). As discussed below, these tuning tasks are very effective in enhancing the alignment between language models and collaborative semantics.

1) *Sequential Item Prediction*: Since our approach is built in a LLM-based generative manner, we consider employing sequential item prediction as the major tuning objective. Specifically, we construct personalized recommendation instructions based on the user's current historical interactions. Then, LLMs are prompted by the instructions and the interaction history, to predict the next item that the target user is likely to interact with. Here, the user's historical interactions are described and identified as an index sequence of interacted items arranged in chronological order. A sample instance is given as follows:

Instruction:

Here are the user's historical interactions: $\langle a_{-124} \rangle \langle b_{-192} \rangle \langle c_{-41} \rangle \langle d_{-17} \rangle, \dots, \langle a_{-82} \rangle \langle b_{-59} \rangle \langle c_{-191} \rangle \langle d_{-66} \rangle$, try to recommend another item to the user. Note that the historical interactions are arranged in chronological order.

Response:

$\langle a_{-112} \rangle \langle b_{-32} \rangle \langle c_{-5} \rangle \langle d_{-175} \rangle$

However, due to the large semantic gap, simply fine-tuning LLMs with the above target task, it is difficult to sufficiently integrate language and collaborative semantics in LLMs.

2) *Explicit Index-Language Alignment*: Although our item indices are constructed based on titles and descriptions of items, they rely solely on shared prefix codewords to establish a weak correlation among items with similar language semantics. To further endow item indices with language semantics, we propose two explicit index-language alignment tasks for tuning LLMs.

On the one hand, the LLM should be capable of accurately identifying the item indices based on the associated title or description. On the other hand, it is expected that LLM can naturally capture relevant item information from its indices. Considering the two aspects, we first instruct the LLM to generate the corresponding item indices according to the item's title/description or a combination of both. Then, we instruct the LLM to recover the item information based on its indices. We

present two instruction samples to illustrate the two alignment tuning tasks in the following.

Instruction:

An item is called “*Pokémon Moon - Nintendo 3DS*” and described as “*Pokémon Moon will launch in the ... become a Pokémon Champion!*”, can you tell me which item it is?

Response:

<a_66><b_197><c_236><d_223>

Instruction:

Please tell me what item <a_66><b_197><c_236><d_223> is called, along with a brief description of it.

Response:

Item Title: *Pokémon Moon - Nintendo 3DS*

Item Description: *Pokémon Moon will launch in the ... become a Pokémon Champion!*

Actually, such a mutual prediction method is essentially similar to what has been in cross-modal semantic alignment. It can be an analogy with the mutual association or mapping between images and text [31], or the conversion between speech and text [53]. By instruction tuning the LLM with these alignment instructions, item indices can be seamlessly integrated into the semantic space spanned by the LLM.

3) *Implicit Recommendation-oriented Alignment*: After being tuned with the above alignment tasks, LLMs can acquire basic knowledge of collaborative semantics. In this part, we further consider enhancing the model capacity via recommendation-oriented alignment tasks, so that LLM can better leverage both language and collaborative semantics to fulfill various recommendation tasks in a more accurate way. Specifically, we design the following three alignment tasks:

a) *Asymmetric item prediction*: As discussed in Section III-C1, for sequential item prediction, both the interaction history (*condition*) and the target item (*target*) are formatted in the representation of item indices. We call this tuning task *symmetric* since both the condition and target for prediction are based on item indices. To further enhance the semantic alignment, we increase the prediction difficulty by changing the representations of condition and target, so as to derive different combinations of semantic representations for items. Specially, we consider the following three representation methods: (1) replacing the indices of target item with the item title, instructing the LLM to generate the item title directly based on the item index sequence; (2) replacing the indices of target item with the item description, instructing the LLM to generate the item features and attributes expected by the user; (3) representing the user interaction history as a text sequence of item titles instead of an index sequence, instructing the LLM to infer user preferences based on the title sequence. The sample instructions for the three scenarios are as follows:

Instruction:

Based on the user’s historical interactions: <a_38><b_94><c_198><d_59>, ..., <a_190><b_60><c_94><d_86>, try to predict the title of the item that the user may need next.

Response:

NBA 2K16 - PlayStation 4

Instruction:

Here is the item interaction history of the user: <a_38><b_94><c_198><d_59>, ..., <a_190><b_60><c_94><d_86>, please tell me what features he expects from his next item.

Response:

The NBA 2K franchise is ... cover of your choice into the front of box.

Instruction:

Given the title sequence of user historical interactive items: “*The Biggest Loser Ultimate Workout - Xbox 360*”, ..., “*Lego: Marvel Super Heroes, XBOX 360*”, can you recommend a suitable next item for the user?

Response:

<a_27><b_58><c_138><d_201>
(*Lego Star Wars - Xbox 360*)

To make a comparison, the item prediction task in Section III-C1 involves mapping an index sequence to the target indices, while the tuning tasks in Section III-C2 explicitly align item indices with their corresponding language information. These asymmetric tasks are more difficult, which enforces LLMs to unify item indices, language semantics, and collaborative semantics for fulfilling the recommendation tasks. As will be shown in the experiment part (Section IV-C), these tuning tasks are useful in adapting LLMs to recommender systems.

b) *Item prediction based on user intention*: Drawing inspiration from [19], a recommender system in real life should possess the ability to understand the actual intentions of users and provide high-quality recommendations accordingly. This leads to a task similar to item retrieval. Referring to the approach in [19], as reviews offer valuable evidence regarding users’ personal tastes and motivations for making a specific interaction, we consider extracting intentions from the related reviews of the target item. To accomplish this, we utilize GPT-3.5 to process these reviews and extract user intentions. As for instructions, we mainly design two types of tasks: the former queries an item recommendation directly based on instant user intention, and the latter provides the user’s interaction history for a personalized recommendation.

Instruction:

Suppose you are a search engine, now a user searches that: “*The game has an open world environment ... activities to complete*”, can you select an item to respond to the user’s query?

Response:

<a_104><b_4><c_47><d_182>
(Grand Theft Auto Vice City Stories)

Instruction:

As a recommender system, you are assisting a user who has recently interacted with the following items: <a_64><b_159><c_1><d_89>, ..., <a_119><b_98><c_162><d_155>. The user expresses a desire to obtain another item with the following characteristics: “*The console offers 500GB of storage, ... 4K HDR gaming*”. Please recommend an item that meets these criteria.

Response:

<a_227><b_206><c_156><d_156>
(PlayStation 4 500GB Console)

c) *Personalized preference inference*: Intuitively, a user’s interaction history can implicitly reflect his or her personal preferences, but explicit preferences are generally absent from the dataset. Thus, we employ GPT-3.5 to infer the user’s explicit preferences from items the user has interacted with in the past. Unlike prior work [19], we infer user preferences based on the index sequence of historical items rather than the title sequence. This task requires the index sequence to act as an effective substitute for the title sequence, enabling the LLM to understand the joint language and collaborative semantics within the index sequence and accurately extract user preferences. The instruction could be:

Instruction:

Utilizing the ordered list of the user’s historical interaction items as a reference, please make an informed estimation of the user’s preferences. The historical interactions are as follows:
<a_227><b_186><c_52><d_166>, ..., <a_120><b_208><c_146><d_153>.

Response:

The user has recently been playing a game that ... with more data storage and/or higher data processing capabilities.

In this work, we mainly focus on the setting of sequential item prediction, *i.e.*, sequential recommendation, while our approach can be easily extended to other tuning tasks in recommender systems, *e.g.*, bundle prediction and explanation generation. Actually, our index mechanism can support various instruction tuning tasks as in standard language models [54], since these indices are endowed with both language and collaborative semantics, acting as common tokens for LLMs.

D. Training and Inference

In this section, we discuss how to optimize our base LLM using the aforementioned tuning tasks and how to utilize it to fulfill the recommendation task over the entire item set.

1) *Training*: In this paper, we strive to leverage the semantic understanding and generation capabilities of LLMs to facilitate sequential recommendation. To this end, we employ LLaMA [25] as our backbone model and then optimize it via instruction tuning. The tuning tasks mentioned above can be conveniently formatted as conditional language generation tasks in a sequence-to-sequence manner. We optimize the negative log-likelihood of the generation target as follows:

$$\mathcal{L} = - \sum_{\langle I, Y \rangle \in \mathcal{B}} \sum_{j=1}^{|Y|} \log P(Y_j | I, Y_{<j}), \quad (7)$$

where $\langle I, Y \rangle$ represents a pair of instruction and target response, Y_j is the j -th token of Y and $Y_{<j}$ denotes the tokens before Y_j . For each task, we designed multiple instruction templates to enhance the instruction diversity. However, during a training epoch, each data is only combined with one sampled instruction template, which is different from those in prior approaches [19], [22], [23]. This strategy is based on our observation that LLaMA, as an LLM with over 7B parameters, achieves better results by examining specific data only a few times [55]. In contrast, repeating data may lead to overfitting.

2) *Inference*: Our objective is to generate the top n items from the entire item set which most match the preference of a given user during inference. To accomplish this, the decoder module performs a beam search across the index tokens. Here, we use the index structure built in Section III-B for item decoding. Additionally, when calculating logits, the probabilities of tokens that may result in illegal item indices will be assigned as 0 to ensure generation quality.

Given an input sequence, the inference time is mainly consumed in the multi-layer self-attention calculation. The time complexity of a forward process in the vanilla Transformer is $\mathcal{O}(N^2dL)$, where L is the number of model layers, N is the sequence length, and d is the dimension of hidden states. Overall, to autoregressively generate complete target item indices, the time consumption is $\mathcal{O}(HN^2dL)$, where H is the number of indices (usually a small value like $H = 4$). But in fact, the key and value tensors of each layer can be cached for subsequent decoding, called *KV Cache* [56]. After applying KV Cache, the complexity can be optimized to $\mathcal{O}(N^2dL + HN^2d)$. In addition to the inference speed, the memory efficiency can also be improved through various technologies such as model quantization [57] and PagedAttention [58].

E. Discussion

In this part, we compare LC-Rec with existing LLM-based methods to highlight the contributions of our approach.

Text-based methods such as TALLRec [21] and InstrucRec [19] typically represent user historical behavior as a sequence of item titles, thereby formatting the sequential recommendation task into a natural language instruction, which

TABLE I

COMPARISON OF LC-REC WITH SEVERAL RELATED STUDIES. “FR” DENOTES FULL RANKING ACROSS THE ENTIRE ITEM SET. “LS” DENOTES LANGUAGE SEMANTICS. “CS” DENOTES COLLABORATIVE SEMANTICS. “ILC” DENOTES THE INTEGRATION OF LANGUAGE AND COLLABORATIVE SEMANTICS IN LLMs.

Methods	Scale	Backbone	FR	LS	CS	ILC
TIGER [45]	N/A	N/A	✓	✗	✓	✗
P5 [22], [23]	220M	T5	✓	✗	✓	✗
InstructRec [19]	3B	Flan-T5	✗	✓	✗	✗
TALLRec [21]	7B	LLaMA	✗	✓	✗	✗
LC-Rec	7B	LLaMA	✓	✓	✓	✓

can be easily adapted to the LLM. However, these methods are not suitable for the full ranking setting, since they have difficulty in generating the item information over the entire item set. They either consider a discrimination question that can be answered with “Yes/No” [21] or perform reranking based on a small number of candidate items [19]. Furthermore, this approach mainly relies on language semantics to tackle the recommendation tasks, which neglects the collaborative semantic information in recommender systems.

Index-based methods such as TIGER [45] and P5 [22], [23] (specifically, we focus on the sequential recommendation task of P5) directly convert the traditional item ID-based sequential recommendation into a generative paradigm. TIGER does not use LLM as a recommender. Instead, it trains a Transformer model from scratch to predict the next item given an input item sequence, where each item is identified by multiple discrete IDs. P5 adapts recommendation tasks into the text-to-text format to enable unified modeling. However, within this framework, sequential recommendation is still organized as a mapping process from item ID sequence to target item ID, which only establishes collaborative semantics between item IDs and is independent of language semantics in LLMs. We are also aware of several concurrent studies [42]–[44] that aim to adapt LLMs for recommender systems. They mainly consider enhancing the semantics of items from different aspects, including setting multi-type identifiers (*e.g.*, ID, title, and attributes) [42], incorporating external collaborative representations [43], and learning dual user/item embeddings [44].

As a comparison, our work focuses on the integration of language and collaborative semantics for enhancing the recommendation capacity of LLMs. Specifically, we adopt a new item indexing mechanism that ensures index uniqueness and effectively reduces the vocabulary size. Moreover, we further design various alignment tasks for enhancing the semantic integration. Based on these improvements, our approach can effectively integrate collaborative semantics into LLMs, and further leverage the enhanced capacity of LLMs for fulfilling the recommendation tasks. The comparison of our method with several related studies is shown in Table I.

IV. EXPERIMENT

In this section, we first set up the experiments and then present the results as well as analyses of our proposed approach.

TABLE II

STATISTICS OF THE PREPROCESSED DATASETS. “Avg. len” REPRESENTS THE AVERAGE LENGTH OF ITEM SEQUENCES.

Datasets	#Users	#Items	#Interactions	Sparsity	Avg. len
Instruments	24,773	9,923	206,153	99.92%	8.32
Arts	45,142	20,957	390,832	99.96%	8.66
Games	50,547	16,860	452,989	99.95%	8.96
CDs	93,652	64,032	1,178,449	99.98%	12.58

A. Experiment Setup

1) *Dataset*: We evaluated the proposed approach on four subsets of Amazon review data [59], including “*Musical Instruments*”, “*Arts, Crafts and Sewing*”, “*Video Games*”, and “*CDs and Vinyl*”. All four datasets contain user review data from May 1996 to October 2018. Each item in the dataset is associated with a title and a description. Following previous work [12], we first filter out unpopular users and items with less than five interactions. Then, we create user behavior sequences based on the chronological order. The maximum item sequence length is uniformly set to 20 to meet all baseline requirements. The statistics of our preprocessed datasets are shown in Table II.

2) *Baseline Models*: We adopt the following representative methods for comparison with our LC-Rec:

- **Caser** [4] is a CNN-based method that models user behaviors by horizontal and vertical convolutional filters.
- **HGN** [26] uses hierarchical gating networks to capture long-term and short-term interests from user behaviors.
- **GRU4Rec** [1] is an RNN-based sequential recommendation model that utilizes GRU to encode the item sequence.
- **BERT4Rec** [8] adopts a bidirectional Transformer model and combines it with a mask prediction task for the modeling of item sequences.
- **SASRec** [2] exploits a unidirectional Transformer to model item sequences and predict the next item.
- **FMLP-Rec** [60] proposes an all-MLP model with learnable filters, which ensures efficiency and reduces noise.
- **FDSA** [9] focuses on the transformation patterns between item features, modeling both item-level and feature-level sequences separately through self-attention networks.
- **S³-Rec** [10] utilizes mutual information maximization to pre-train a sequential recommendation model, learning the correlation between items and attributes.
- **P5-CID** [22], [23] organizes multiple recommendation tasks in a text-to-text format and models different tasks uniformly using the T5 model. Subsequently, the author team explores the construction of item indices for sequential recommendation, including sequential indexing and collaborative indexing. Here, we employ P5 with collaborative indexing as the baseline and implement it according to the code² provided by the authors.
- **TIGER** [45] adopts the generative retrieval paradigm for sequential recommendation and introduces a semantic ID to uniquely identify items. Due to the official code

²<https://github.com/Wenyueh/LLM-RecSys-ID/>

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT METHODS ON THE FOUR DATASETS. THE BEST AND SECOND-BEST PERFORMANCES ARE INDICATED IN BOLD AND UNDERLINED FONT, RESPECTIVELY. THE PERFORMANCE FOR OUR LC-REC IS AVERAGE RESULTS FROM MULTIPLE INSTRUCTION TEMPLATES.

Dataset	Metrics	Caser	HGN	GRU4Rec	BERT4Rec	SASRec	FMLP-Rec	FDSA	S ³ -Rec	P5-CID	TIGER	LlamaRec	LC-Rec	Improv.
Instruments	HR@1	0.0149	0.0523	0.0571	0.0435	0.0503	0.0480	0.0520	0.0367	0.0587	<u>0.0608</u>	0.0505	0.0706	+16.12%
	HR@5	0.0543	0.0813	0.0821	0.0671	0.0751	0.0786	0.0834	<u>0.0863</u>	0.0827	<u>0.0863</u>	0.0855	0.1002	+16.11%
	HR@10	0.0710	0.1048	0.1031	0.0822	0.0947	0.0988	0.1046	<u>0.1136</u>	0.1016	0.1064	0.1110	0.1220	+7.39%
	NDCG@5	0.0355	0.0668	0.0698	0.0560	0.0627	0.0638	0.0681	0.0626	0.0708	<u>0.0738</u>	0.0678	0.0856	+15.99%
	NDCG@10	0.0409	0.0744	0.0765	0.0608	0.0690	0.0704	0.0750	0.0714	0.0768	<u>0.0803</u>	0.0757	0.0926	+15.32%
Arts	HR@1	0.0138	0.0300	0.0421	0.0337	0.0225	0.0310	0.0451	0.0245	<u>0.0485</u>	0.0138	0.0428	0.0634	+30.72%
	HR@5	0.0379	0.0622	0.0749	0.0559	0.0757	0.0757	0.0734	0.0767	0.0724	0.0788	<u>0.0854</u>	0.1011	+18.38%
	HR@10	0.0541	0.0875	0.0964	0.0713	0.1016	0.1046	0.0933	0.1051	0.0902	0.1012	<u>0.1109</u>	0.1266	+14.16%
	NDCG@5	0.0262	0.0462	0.0590	0.0451	0.0508	0.0541	0.0595	0.0521	0.0607	0.0631	<u>0.0647</u>	0.0828	+27.98%
	NDCG@10	0.0313	0.0544	0.0659	0.0500	0.0592	0.0634	0.0660	0.0612	0.0664	0.0703	<u>0.0726</u>	0.0906	+24.79%
Games	HR@1	0.0085	0.0154	0.0176	0.0136	0.0145	0.0152	0.0161	0.0119	0.0177	<u>0.0188</u>	0.0148	0.0317	+68.62%
	HR@5	0.0367	0.0517	0.0586	0.0482	0.0581	0.0571	0.0644	0.0606	0.0506	0.0599	<u>0.0652</u>	0.0800	+22.70%
	HR@10	0.0617	0.0856	0.0964	0.0763	0.0940	0.0930	0.1041	0.1002	0.0803	0.0939	<u>0.1092</u>	0.1174	+7.51%
	NDCG@5	0.0227	0.0333	0.0381	0.0311	0.0365	0.0361	0.0404	0.0364	0.0342	0.0392	<u>0.0411</u>	0.0560	+36.25%
	NDCG@10	0.0307	0.0442	0.0502	0.0401	0.0481	0.0476	0.0531	0.0491	0.0437	0.0501	<u>0.0545</u>	0.0681	+24.95%
CDs	HR@1	0.0063	0.0114	0.0265	0.0212	0.0186	0.0231	0.0275	0.0203	<u>0.0302</u>	0.0269	0.0281	0.0404	+33.77%
	HR@5	0.0249	0.0404	0.0612	0.0529	0.0654	0.0645	0.0662	0.0677	0.0597	0.0511	<u>0.0686</u>	0.0769	+12.10%
	HR@10	0.0419	0.0653	0.0868	0.0768	0.0957	<u>0.1005</u>	0.0962	0.1004	0.0806	0.0707	0.0980	0.1092	+8.66%
	NDCG@5	0.0154	0.0256	0.0450	0.0372	0.0425	0.0435	0.0469	0.0443	0.0460	0.0390	<u>0.0480</u>	0.0588	+22.50%
	NDCG@10	0.0209	0.0336	0.0532	0.0448	0.0522	0.0551	0.0566	0.0549	0.0520	0.0453	<u>0.0576</u>	0.0664	+15.28%

not being released by the authors, here we implement it ourselves by Transformers³ following the implementation details provided in the paper.

- **LlamaRec** [39] proposes a two-stage framework, which first performs ID-based retrieve, and then uses fine-tuned LLaMA-2 [25] for reranking to improve performance.

3) *Evaluation Settings*: To evaluate the performance of sequential recommendation, we adopt two widely used metrics, top- K Hit Ratio (HR) and top- K Normalized Discounted Cumulative Gain (NDCG). In this paper, we set K as 1, 5, and 10. Following previous works [2], [10], [60], we employ the *leave-one-out* strategy for evaluation. Concretely, for each user behavior sequence, the most recent item is used as the test data, the second most recent item is used as the validation data, and the remaining interaction records are used for training. We perform full ranking evaluation over the entire item set instead of sample-based evaluation. For the generative methods based on beam search, the beam size is uniformly set to 20.

4) *Implementation Details*: To construct item indices, we utilize LLaMA to encode the title and description of the item as its embedding and use mean pooling to aggregate representations. The level of item indices is set to 4, with each level consisting of 256 codebook vectors, and vector dimension is 32. Both the encoder and decoder of RQ-VAE are implemented as Multi-Layer Perceptrons (MLPs). The model is optimized by the AdamW optimizer, employing a learning rate of 0.001 and a batch size of 1024.

For LLM fine-tuning, we implemented LC-Rec based on LLaMA through Transformers³ and accelerated training by DeepSpeed⁴. All tokens related to item indices are appended to

the tokenizer as out-of-vocabulary (OOV) tokens. We employ the AdamW optimizer for model optimization, setting the learning rate to 5e-5 and weight decay to 0.01. During the fine-tuning, a cosine scheduler with warmup is utilized to adjust the learning rate. With the application of data parallelism and gradient accumulation, the overall batch size amounts to 128. We conduct training for 4 epochs on each dataset. To prevent overfitting, we ensure that during an epoch, each data is combined with one sampled instruction template and appears only once.

B. Overall Performance

We compare LC-Rec with the different baseline models on four datasets, and the overall results are shown in Table III. Based on these results, we can find:

For the baseline methods, the sequential recommendation methods that incorporate item content information (*i.e.*, FDSA and S³-Rec) perform better than traditional sequential recommendation methods that solely rely on ID and collaborative relationships (*i.e.*, Caser, HGN, GRU4Rec, BERT4Rec, SASRec, FMLP-Rec) on several datasets. This indicates that item content introduced as additional information can effectively improve recommendation performance. As for P5-CID and TIGER, they demonstrate competitive performance across the first two datasets, particularly excelling in HR@1 and the metrics related to item ranking (*i.e.*, NDCG). In terms of Games and CDs datasets, there is no significant improvement compared to traditional sequential recommendation models. One possible reason is that the effectiveness of content information and the challenges faced vary across different data and scenarios. Furthermore, LlamaRec is very competitive compared to traditional models, but it is constrained by the retrieval process, with limited room for improvement.

³<https://github.com/huggingface/transformers/>

⁴<https://github.com/microsoft/DeepSpeed/>

TABLE IV
ABLATION STUDY OF VARIOUS SEMANTIC ALIGNMENT TASKS IN LC-REC. WE SHOW THE RESULTS ON ARTS AND GAMES DATASET.

Methods	Arts					Games				
	HR@1	HR@5	HR@10	NDCG@5	NDCG@10	HR@1	HR@5	HR@10	NDCG@5	NDCG@10
SEQ	0.0561	0.0909	0.1133	0.0740	0.0812	0.0243	0.0626	0.0930	0.0437	0.0535
+ MUT	0.0593	0.0926	0.1141	0.0765	0.0832	0.0275	0.0703	0.1038	0.0491	0.0598
+ ASY	0.0602	0.0945	0.1172	0.0776	0.0848	0.0281	0.0725	0.1073	0.0506	0.0615
+ ITE	0.0638	0.0996	0.1232	0.0813	0.0889	0.0294	0.0770	0.1125	0.0534	0.0648
+ PER	0.0634	0.1011	0.1266	0.0828	0.0906	0.0317	0.0800	0.1174	0.0560	0.0681

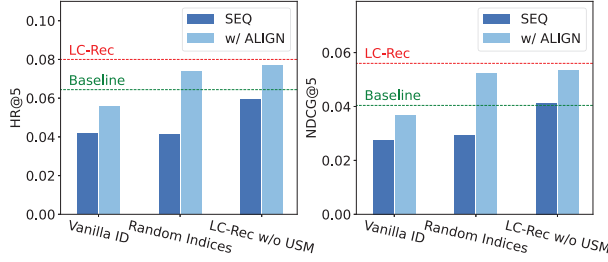


Fig. 2. The performance of our three indexing methods on Games dataset. “SEQ” denotes fine-tuning only with the sequential item prediction task. “w/ ALIGN” denotes combining with our semantic alignment tasks.

Our proposed LC-Rec consistently maintains the best performance on four datasets and shows significant improvements compared to the baseline methods. This superior performance can be attributed to two factors: (1) The item indexing mechanism via vector quantization combined with uniform semantic mapping, which captures similarities between items and ensures a semantically lossless generation process at the last index level. (2) The effective integration of collaborative semantics into LLMs, which results in a seamless fusion of language semantics and collaborative semantics. By employing these strategies, our approach is able to leverage the powerful modeling capabilities of LLMs, thereby achieving significant improvements in the recommendation task.

C. Ablation Study

a) *Various semantic alignment tasks:* Our proposed LC-Rec consists of various semantic alignment tasks, including (1) SEQ: the sequential item prediction task introduced in Section III-C1 as our primary objective, (2) MUT: the mutual prediction task for explicit index-language alignment in Section III-C2, (3) ASY: the asymmetric item prediction task in Section III-C3a, (4) ITE: the item prediction based on user intention in Section III-C3b, (5) PER: the personalized preference inference task in Section III-C3c. The latter three tasks all belong to the implicit recommendation-oriented alignment introduced in Section III-C3. To validate the effectiveness of each component, we conduct the ablation study on Arts and Games dataset to analyze the contribution of each part. The results, as shown in Table IV, indicate that the gradual incorporation of multiple semantic alignment tasks into sequential recommendation, which involves only collaborative semantics, can significantly improve performance. All these

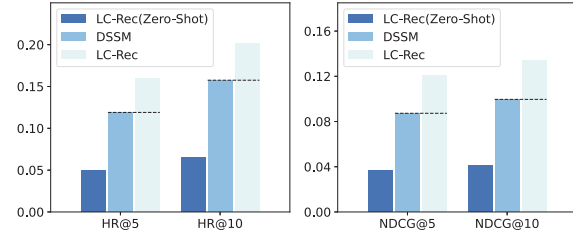


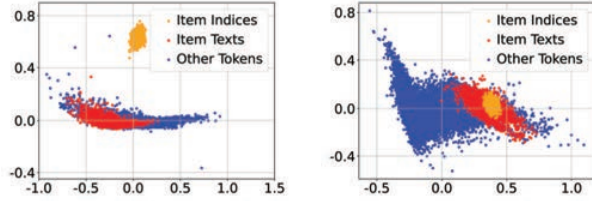
Fig. 3. Performance of item prediction based on user intention.

instruction tuning tasks in LC-Rec are beneficial to enhance sequential recommendation, and there is potential for further improvements by adding more semantic alignment tasks.

b) *Other item indexing methods:* In addition to the semantic alignment tasks, we also examine the proposed item indexing method, by comparing it to another three indexing methods. (1) Vanilla ID is the same as the traditional recommendation model, using a single and unique ID for each item. (2) Random Indices uses multi-level indexing, but the indices at each level are derived from random sampling and are not semantically related. (3) LC-Rec w/o USM removes the uniform semantic mapping in our indexing method and assigns a distinct supplementary index ID to each conflicting item. As shown in Figure 2, our approach (red dotted line, LC-Rec) outperforms all three base indexing methods, indicating the effectiveness of the proposed item indexing method (Section III-B). In addition, if we apply the proposed semantic alignment tasks (“w/ ALIGN” in Figure 2) to these three base indexing methods, their performance can be boosted by a large margin, especially for methods also based on multi-level indexing (e.g., Random Indices and LC-Rec w/o USM), outperforming all baseline methods. The results also demonstrate that the proposed alignment tasks can improve recommendation performance in an indexing-agnostic way.

D. Further Analysis

a) *Item prediction based on user intention:* We further evaluate the ability of LC-Rec to understand the semantics contained in item indices. The evaluation is performed through a user intention-based item prediction task on Games dataset, as described in Section III-C3b. Following the widely used setups in sequential recommendation, the most recent record in each user behavior sequence is used for testing. User intentions are used as the query and are generated by GPT-



(a) Fine-tuning only with the target sequential item prediction task. (b) LC-Rec that consists of a series of alignment tasks.

Fig. 4. 2D visualization of LLM token embeddings via PCA.

3.5 based on review data. We employ DSSM [61], a widely validated retrieval model, as our baseline. It adopts a two-tower architecture to search for relevant items based on textual similarity between a given user query and item titles. In our implementation, BERT [62] is used to encode queries and item titles. As shown in Figure 3, our approach exhibits a significant performance improvement compared to the baseline model. This improvement can be attributed to the integration of language and collaborative semantics in the LLM through item indices. Additionally, “LC-Rec (Zero-Shot)” represents the LC-Rec variant that is not trained in the item prediction task regarding user intention. Interestingly, we can observe that basic language and collaborative semantic alignment can still link item indices to user intentions to some extent, even without prior training on the target task.

b) Embedding visualization analysis: To further investigate the effects of our proposed framework in terms of semantic integration, we delve deeper to study the connection between item indices and the original semantic space of the LLM. Following previous work [63], [64], we employ Principal Component Analysis (PCA) to visualize the embeddings corresponding to different tokens. As shown in Figure 4, “Item Indices” represents index tokens added to the vocabulary, while “Item Texts” represents tokens related to item texts (e.g., title and description). According to 2D visualization results, it is evident that a lack of semantic integration leads to incompatibility between item index tokens and the LLM semantic space. In contrast, our framework is capable of incorporating item indices into the LLM and aligning language and collaborative semantics.

c) Performance on semantically similar negative items: In order to comprehend why integrating language and collaborative semantics can improve LLMs in recommendation tasks, we further evaluate our LC-Rec with a ranking task with different negatives that are similar to ground truth in either language or collaborative semantics. Specifically, we first select two types of semantically similar negative items: (1) Items with similar language semantics, which are selected based on the cosine similarity between item text embeddings. (2) Items with similar collaborative semantics, which are selected based on the cosine similarity between item embeddings from the trained SASRec [2] model. Subsequently, we utilize the model

TABLE V
PERFORMANCE ON SEMANTICALLY SIMILAR NEGATIVE ITEMS.

Methods	Language Neg.	Collaborative Neg.	Random Neg.
SASRec	73.52	52.25	89.78
LLaMA	56.67	51.23	61.14
ChatGPT	60.94	51.30	66.66
LC-Rec (Title)	67.74	56.72	84.64
LC-Rec	75.73	60.01	90.19

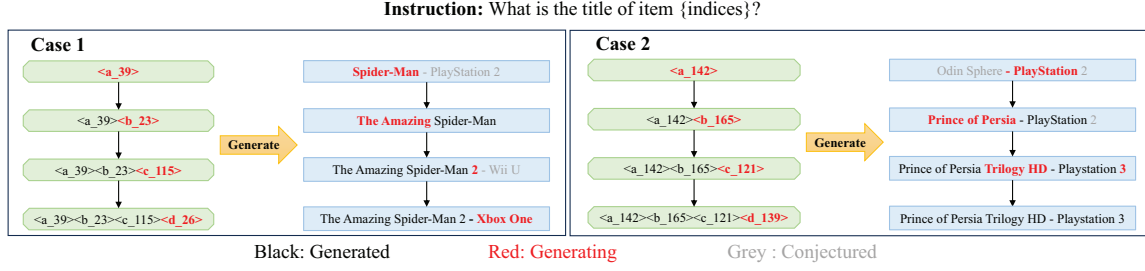
to choose between the ground-truth target item and the negative item with similar language/collaborative semantics. In addition, we use random negative items as a comparison benchmark and measure the performance by accuracy.

We adopt SASRec, LLaMA without fine-tuning, and ChatGPT as the comparison methods. “LC-Rec (Title)” refers to our approach but makes recommendations based on item titles rather than indices. The results are shown in Table V. In the task of distinguishing items with similar language semantics, our method achieved the best performance, benefiting from the integration of collaborative semantics implied by recommender systems. Additionally, substituting item indices with titles for recommendations also yielded competitive results, owing to implicit alignment between item indices and titles within our model. Another task, distinguishing items with similar collaborative semantics, is often considered more challenging. This is due to the fact that the item with similar collaborative semantics may also have language semantic relevance to the ground-truth target item. However, even for such a difficult task, our LC-Rec still shows better performance than the strong baselines, thanks to the unification of language and collaborative semantics. Furthermore, the non-fine-tuned LLaMA and ChatGPT perform sub-optimally in these challenging scenarios, demonstrating that utilizing LLMs directly for recommendation purposes is often inadequate due to the large gap between recommendation tasks and natural language tasks.

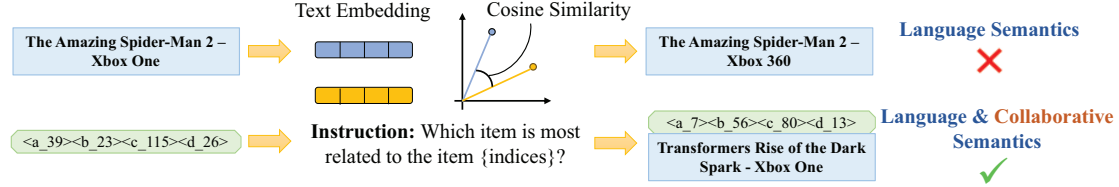
E. Case Study

To intuitively explore the semantic information implicitly learned in the item indices, we present two types of illustrative cases in Figure 5.

On the one hand, we analyze the hierarchical semantics in the multi-level item index. Specifically, we initially attempt to generate the item title using only the first index and gradually include more until all four indices are used. As shown in Figure 5(a), when relying solely on the first-level index, the generated content often fails to match the ground-truth item, but it already possesses some relevant semantic information. For example, in the first case, a single index can generate the keyword “Spider-Man”, whereas in the second case, a game belonging to the same categories (i.e., adventure) and similar platform (i.e., PlayStation) as the ground-truth item can be generated. As more indices are included, the generated content progressively converges towards the target title. Notably, at the second level, our LC-Rec is already capable of inferring the item name to a significant extent. The subsequent third



(a) Generate the item title based on different number of indices.



(b) Related items are generated based on item indices or recalled based on text embedding cosine similarity.

Fig. 5. Case study about the semantics within item indices. For the cases in Figure 5(a), it can be observed that as the number of index increases, the generated content progressively converges towards the target title, and the semantic changes show a trend from coarse to fine. For the cases in Figure 5(b), compared to those based solely on language semantics, related items generated using item indices that integrate both language and collaborative semantics are more suitable for recommendation scenarios.

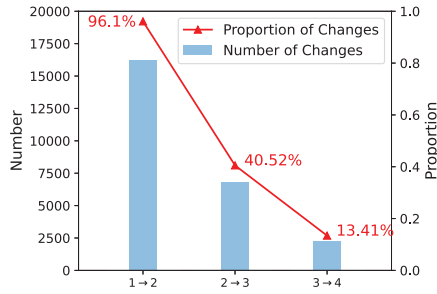


Fig. 6. Content changes caused by each level index.

level further refines the semantic information, while the fourth level contains relatively less semantic information, which is consistent with the coarse-to-fine quantization process employed during index construction. Moreover, we also count the proportion of generated results changes caused by each level of indices. As shown in Figure 6, also consistent with our conjecture, the proportion of content changes gradually decreases as the index level increases.

On the other hand, we try to generate the item that are most relevant or similar to a given item through its indices. We then compare the generated results with the similar item obtained based on cosine similarity between item text embeddings. As presented in Figure 5(b), the similar item generated by our LC-Rec is a game of the same category and platform as the source item, while a duplicate game for another platform is obtained simply based on language semantic similarity. In recommendation scenarios, the former that integrates both

language and collaborative semantics is usually more suitable to meet user needs.

V. CONCLUSION

In this paper, we proposed a LLM-based recommendation approach, named LC-Rec. In order to adapt LLMs to sequential recommendation tasks, we focused on two main aspects: item indexing and alignment tuning. Concretely, we introduced a vector quantization method combined with uniform semantic mapping for item index learning. To facilitate the integration of item indices into the LLM, we proposed a series of semantic alignment tasks to align language and collaborative semantics for recommendation. These tasks include sequential item prediction, explicit index-language alignment, and implicit recommendation-oriented alignment. Based on the learned item indices, our approach employed these alignment tuning tasks to effectively adapt LLMs for sequential recommendation. Extensive experiments on three large datasets demonstrated the effectiveness of our approach, outperforming a number of competitive baseline models.

As future work, we will explore how to extend the current approach in a multi-turn chat setting, so that it can support more flexible interaction with users. In addition, we will also investigate how to better reserve the general abilities of LLMs when making domain adaptations.

ACKNOWLEDGMENT

This work was partially supported by National Natural Science Foundation of China under Grant No. 62222215 and Beijing Natural Science Foundation under Grant No. 4222027. Xin Zhao is the corresponding author.

REFERENCES

- [1] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.
- [2] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *ICDM*, 2018.
- [3] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *CIKM*, 2017.
- [4] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *WSDM*, 2018.
- [5] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," in *WSDM*, 2019.
- [6] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *AAAI*, 2019.
- [7] C. Xu, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, F. Zhuang, J. Fang, and X. Zhou, "Graph contextualized self-attention network for session-based recommendation," in *IJCAI*, 2019.
- [8] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *CIKM*, 2019.
- [9] T. Zhang, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, D. Wang, G. Liu, X. Zhou *et al.*, "Feature-level deeper self-attention network for sequential recommendation," in *IJCAI*, 2019.
- [10] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *CIKM*, 2020.
- [11] Y. Xie, P. Zhou, and S. Kim, "Decoupled side information fusion for sequential recommendation," in *SIGIR*, 2022.
- [12] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen, "Towards universal sequence representation learning for recommender systems," in *SIGKDD*, 2022.
- [13] J. Li, M. Wang, J. Li, J. Fu, X. Shen, J. Shang, and J. McAuley, "Text is all you need: Learning language representations for sequential recommendation," in *SIGKDD*, 2023.
- [14] H. Ding, Y. Ma, A. Deoras, Y. Wang, and H. Wang, "Zero-shot recommender systems," *arXiv preprint arXiv:2105.08318*, 2021.
- [15] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," in *WWW*, 2023.
- [16] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [17] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu *et al.*, "A survey on large language models for recommendation," *arXiv preprint arXiv:2305.19860*, 2023.
- [18] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," *arXiv preprint arXiv:2305.08845*, 2023.
- [19] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," *arXiv preprint arXiv:2305.07001*, 2023.
- [20] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, "M6-rec: Generative pretrained language models are open-ended recommender systems," *arXiv preprint arXiv:2205.08084*, 2022.
- [21] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "Tallrec: An effective and efficient tuning framework to align large language model with recommendation," *arXiv preprint arXiv:2305.00447*, 2023.
- [22] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *RecSys*, 2022.
- [23] W. Hua, S. Xu, Y. Ge, and Y. Zhang, "How to index item ids for recommendation foundation models," *SIGIR-AP*, 2023.
- [24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *NeurIPS*, 2022.
- [25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [26] C. Ma, P. Kang, and X. Liu, "Hierarchical gating networks for sequential recommendation," in *SIGKDD*, 2019.
- [27] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *WWW*, 2010.
- [28] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *ICDM*, 2016.
- [29] J. Harte, W. Zörgdrager, P. Louridas, A. Katsifodimos, D. Jannach, and M. Fraggoulis, "Leveraging large language models for sequential recommendation," in *RecSys*, 2023, pp. 1096–1102.
- [30] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *ICLR*, 2023.
- [31] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.
- [32] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen, "Structgpt: A general framework for large language model to reason over structured data," *arXiv preprint arXiv:2305.09645*, 2023.
- [33] A. Acharya, B. Singh, and N. Onoe, "LLM based generation of item-description for recommendation system," in *RecSys*, 2023, pp. 1204–1207.
- [34] D. Di Palma, "Retrieval-augmented recommender system: Enhancing recommender systems with large language models," in *RecSys*, 2023, pp. 1369–1373.
- [35] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chatrec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.
- [36] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, Z. Sun, X. Zhang, and J. Xu, "Uncovering chatgpt's capabilities in recommender systems," in *RecSys*, 2023.
- [37] L. Wang and E.-P. Lim, "Zero-shot next-item recommendation using large pretrained language models," *arXiv preprint arXiv:2304.03153*, 2023.
- [38] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. J. McAuley, "Large language models as zero-shot conversational recommenders," in *CIKM*, 2023, pp. 720–730.
- [39] Z. Yue, S. Rabhi, G. de Souza Pereira Moreira, D. Wang, and E. Oldridge, "Llamarec: Two-stage recommendation using large language models for ranking," *arXiv preprint arXiv:2311.02089*, 2023.
- [40] B. Yin, J. Xie, Y. Qin, Z. Ding, Z. Feng, X. Li, and W. Lin, "Heterogeneous knowledge fusion: A novel approach for personalized recommendation via LLM," in *RecSys*, 2023, pp. 599–601.
- [41] L. Li, Y. Zhang, and L. Chen, "Prompt distillation for efficient llm-based recommendation," in *CIKM*, 2023, pp. 1348–1357.
- [42] X. Lin, W. Wang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua, "A multi-facet paradigm to bridge large language model and recommendation," *arXiv preprint arXiv:2310.06491*, 2023.
- [43] Y. Zhang, F. Feng, J. Zhang, K. Bao, Q. Wang, and X. He, "Collm: Integrating collaborative embeddings into large language models for recommendation," *arXiv preprint arXiv:2310.19488*, 2023.
- [44] Y. Zhu, L. Wu, Q. Guo, L. Hong, and J. Li, "Collaborative large language model for recommender systems," *arXiv preprint arXiv:2311.01343*, 2023.
- [45] S. Rajput, N. Mehta, A. Singh, R. H. Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Q. Tran, J. Samost, M. Kula, E. H. Chi, and M. Sathiamoorthy, "Recommender systems with generative retrieval," in *NeurIPS*, 2023.
- [46] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [47] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *CVPR*, 2022.
- [48] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma, "Learning discrete representations via constrained clustering for effective and efficient dense retrieval," in *WSDM*, 2022.
- [49] Y. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, 2019.
- [50] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *NeurIPS*, 2020.
- [51] S. Lin, C. Liu, P. Zhou, Z.-Y. Hu, S. Wang, R. Zhao, Y. Zheng, L. Lin, E. Xing, and X. Liang, "Prototypical graph contrastive learning," *TNNLS*, 2022.
- [52] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *NIPS*, vol. 26, 2013.
- [53] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, B. Chen, C. Lei, A. Liu, C. Song, X. Lei *et al.*, "Unified language-vision pretraining with dynamic discrete visual tokenization," *arXiv preprint arXiv:2309.04669*, 2023.

- [54] X. Wang, X. Tang, X. Zhao, J. Wang, and J. Wen, "Rethinking the evaluation for conversational recommendation in the era of large language models," in *EMNLP*, 2023.
- [55] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," in *ACL*, 2022.
- [56] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, J. Heek, K. Xiao, S. Agrawal, and J. Dean, "Efficiently scaling transformer inference," *MLSys*, vol. 5, 2023.
- [57] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," *arXiv preprint arXiv:2103.13630*, 2021.
- [58] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *SOSP*, 2023.
- [59] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *EMNLP-IJCNLP*, 2019.
- [60] K. Zhou, H. Yu, W. X. Zhao, and J.-R. Wen, "Filter-enhanced mlp is all you need for sequential recommendation," in *WWW*, 2022.
- [61] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *CIKM*, 2013.
- [62] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.
- [63] J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T. Liu, "Representation degeneration problem in training natural language generation models," in *ICLR*, 2018.
- [64] L. Wang, J. Huang, K. Huang, Z. Hu, G. Wang, and Q. Gu, "Improving neural language generation with spectrum control," in *ICLR*, 2019.