RecBase: Generative Foundation Model Pretraining for Zero-Shot Recommendation

Sashuai Zhou[1,3], Weinan Gan[2], Qijiong Liu[4], Ke Lei[1], Jieming Zhu[2†], Hai Huang[1], Yan Xia[1], Ruiming Tang[2], Zhenhua Dong[2], Zhou Zhao[1,3†]
[1]Zhejiang University [2]Huawei Noah's Ark Lab [3]Shanghai AI Lab [4]The HK PolyU
zhousashuai@zju.edu.cn jiemingzhu@ieee.org zhaozhou@zju.edu.cn

Abstract

Recent advances in LLM-based recommendation have shown promise, yet their cross-domain generalization is hindered by a fundamental mismatch between language-centric pretraining and the recommendation task. Existing methods, relying on language-level knowledge, fail to capture dynamic, item-level user interests across domains. To bridge this gap, we propose RecBase, a domain-agnostic foundational model pretrained with a recommendation-oriented objective. RecBase leverages a large-scale, heterogeneous, cross-domain corpus with unified textual representations and feature mappings to enhance cross-domain generalization. To further align item semantics across domains, we introduce a unified item tokenizer that encodes items into hierarchical concept identifiers, enabling structured representation and efficient vocabulary sharing. The model is trained using an autoregressive objective to capture complex item-level sequential patterns. On eight real-world datasets, our 1.5B-parameter model matches or surpasses the performance of LLM baselines up to 7B parameters in zero-shot and cross-domain recommendation tasks.

1 Introduction

In recent years, large language models (LLMs) (Minaee et al., 2024) have demonstrated powerful capabilities in zero-shot learning, multi-task unification, and multi-domain generalization. Inspired by these successes, an intriguing yet underexplored area is the development of foundational models specifically tailored for recommender systems (Huang et al., 2024; Liu et al., 2024b). Recommender systems (Zhu et al., 2022) are essential for helping users discover content of interest and have been widely applied across various domains, such as videos, music, and products. An ideal foundation model for recommender systems should be capable of addressing diverse recommendation tasks across different domains while also performing effectively in zero-shot and few-shot (e.g., cold-start) settings.

Toward this goal, existing research aims to leverage the strengths of LLMs to enhance the effectiveness and versatility of recommender systems, a field referred to as LLM-based recommendation (Wu et al., 2024). For instance, some initial efforts, such as those described in (Liu et al., 2023; Liang et al., 2025; Hou et al., 2024b), explore the direct application of LLMs for zero-shot recommendation. Other works, like P5 (Geng et al., 2022) and GenRec (Ji et al., 2023), focus on continuing pretraining of LLMs in multi-domain and multi-task recommendation settings. Additionally, some other studies such as (Petrov and Macdonald, 2023; Bao et al., 2023; Lin et al., 2024) investigate efficient fine-tuning and alignment of LLMs for downstream recommendation tasks. However, these approaches face several limitations: 1) Input Representation: Recommendation data often needs to be mapped into language modalities, which may not effectively represent user sequences as shown in our experiments. 2) Knowledge Gap: The knowledge gap between language models and recommendation tasks makes them struggle with modeling item-item co-relationships, hindering their performance in zero-shot recommendations. 3) Model Alignment: Fine-tuning language models to align with recommendation models using downstream task datasets can compromise the model's ability to effectively handle zero-shot and cross-domain recommendations. As a result, such approaches often fail to meet the expectations set for foundational recommendation models.

In this paper, we aim to bridge the gap by making the first effort to pretrain a foundational model from scratch (dubbed RecBase), supporting both zero-shot and multi-domain recommendation settings. To achieve this, we leverage LLMs solely as encoders for unified semantic rep-

†Corresponding Authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15587–15599
November 4-9, 2025

resentation and then model item-item relationships through generative pretraining on large-scale, open-domain, recommendation-oriented item sequence data. Specifically, our work makes the following technical contributions: **1) Data Collection and Representation**: We compile a large-scale, open-domain recommendation dataset spanning 15 different domains (comprising 4.5M items and 35M interactions). We uniformly extract textual representations of items to serve as a data source for pretraining across various domains. **2) Unified Item Tokenizer**: Instead of relying on ID-based sequence modeling, which lacks semantics, or language-based modeling, which is often verbose, we propose a general item tokenizer that unifies item representations across domains. Each item is tokenized into multi-level concept IDs, learned in a coarse-to-fine manner inspired by curriculum learning. This hierarchical encoding facilitates semantic alignment, reduces vocabulary size, and enables effective knowledge transfer across diverse domains. **3) Autoregressive Pretraining**: We adopt an autoregressive modeling paradigm for pretraining, where the model predicts the next token in a sequence. This approach enables learning item co-relationships within a unified concept token space, thereby enhancing the model's generalization in zero-shot and cross-domain settings. For clarity and consistency, we provide two pretrained versions of our foundation models: RecBase-0.3B and RecBase-1.5B.

To assess the model's generalization ability in zero-shot and multi-domain settings, we design an evaluation framework that predicts, i.e., ranks, items users are likely to engage with across a wide range of recommendation tasks. We conduct extensive experiments on eight diverse and previously unseen datasets to examine the zero-shot and cross-domain performance of RecBase. The results show that our recommendation-oriented pretraining strategy substantially enhances the model's ability to generalize to new domains without task-specific tuning. While fine-tuning on in-domain data can further improve performance, our primary emphasis in this work is on zero-shot generalization. In these settings, RecBase consistently outperforms language-model-based baselines, underscoring the effectiveness of recommendation-aligned pretraining and its potential to support robust and adaptable recommendation across domains. RecBase will be open-sourced at https://github.com/reczoo/RecBase.

2 Related Work

2.1 LLM-based Recommendation

Recent advancements in LLM-based recommendation systems have focused on both item scoring and generation tasks. In item scoring, models like M6-Rec (Cui et al., 2022), Prompt4NR (Zhang and Wang, 2023), TabLLM (Hegselmann et al., 2023), and TALLRec (Bao et al., 2023) transform user-item data into natural language representations, either generating item descriptions for scoring or reframing the task as a cloze-style prediction. Similarly, ONCE (Liu et al., 2024a) offers a generative framework for content-based recommendation, and CLLM4Rec (Zhu et al., 2024b) enhances collaboration in LLM-based systems. For item generation, approaches like GPT4Rec (Petrov and Macdonald, 2023), P5 (Geng et al., 2022), EAGER (Wang et al., 2024a), and EAGER-LLM (Hong et al., 2025) leverage generative models to predict the next item based on user behavior, with DiffuRec incorporating uncertainty into sequential recommendations. Additionally, GIRL (Zheng et al., 2023) demonstrates how LLMs can improve job recommendations. While LLM-based methods have advanced recommendations, they struggle with the semantic gap between language-based representations and structured recommendation data. Based on this, we propose a pretrained model that directly learns from recommendation-specific representations to bridge the semantic gap and enhance adaptability across domains.

2.2 Item Representation for Recommendation

Recent works have made significant strides in enhancing recommendation systems by improving item representations. Hierarchical models, such as those using graph neural networks to aggregate item information into representations (Li et al., 2020; Wang et al., 2021a), have been shown to refine user profiles and capture dependencies within and across interactions. Techniques like cross-view contrastive learning (Ma et al., 2022) have also been proposed to model user-bundle and user-item interactions, facilitating better generalization across domains. Additionally, methods addressing sequential recommendation challenges have focused on mitigating item representation divergence (Peng et al., 2022), further enhancing learning efficiency. Another noteworthy advancement is the use of Semantic IDs for items (Rajput et al., 2023; Zheng et al., 2024; Wang et al., 2024b; Zhu

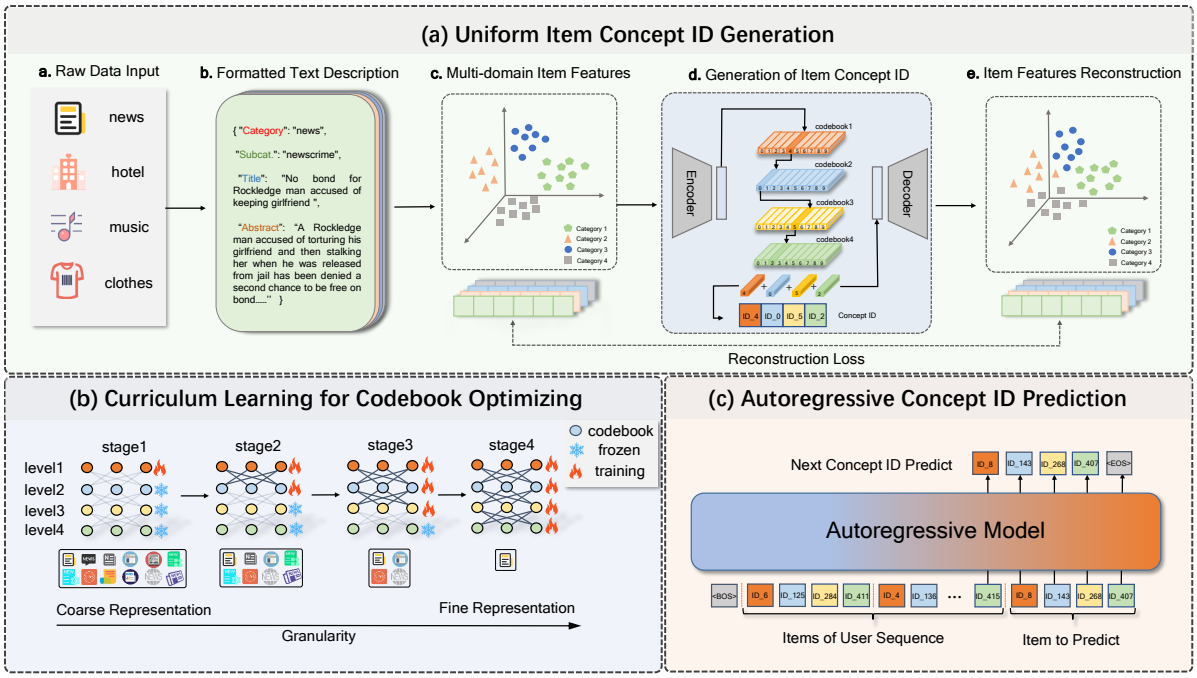


Figure 1: Overview of the RecBase model. (a) illustrates the use of discrete representation techniques to transform product descriptions from multiple domains into unified concept ID sequences. (b) depicts the curriculum learning process for optimizing codebook learning. (c) demonstrates how the autoregressive model leverages discretized concept IDs to predict the next item in the sequence, effectively capturing item relationships for recommendation.

et al., 2024a), where generative retrieval frameworks predict the next item in a sequence, enhancing generalization, particularly in zero-shot scenarios. While these methods show progress, they are often constrained by domain-specific data, limiting their ability to generalize across different contexts. Our approach aims to overcome these limitations by constructing a unified feature representation that enables better generalization across diverse domains.

3 Method

Our proposed method consists of two main stages. First, item representations are mapped into a unified discretized space, generating a concept ID sequence for each item (Section 3.2). Second, this concept ID sequence is used to autoregressively predict the next in the sequence (Section 3.3).

3.1 Preliminary

To facilitate the learning of user behavior based on item history by large language models, it is necessary to discretize continuous item semantic embeddings into unified discrete tokens. This can be achieved through Residual Quantized Variational Autoencoder (RQ-VAE) (Lee et al., 2022). Given an input $e \in \mathbb{R}^d$, RQ-VAE first encodes e into the latent space using an encoder $\mathcal{E}$:

$$z := \mathcal{E}(e) \tag{1}$$

RQ-VAE extends Vector Quantized Variational Autoencoder (VQ-VAE) (van den Oord et al., 2018) by introducing the concept of hierarchical quantization, which progressively quantizes the latent representation z. Specifically, at each level d, the residual r_d is quantized by mapping it to the nearest embedding e_{c_d} in the level-specific codebook $C_d := e_k{}_{k=1}^{K}$, where

$$c_d = \arg\min_k \|r_d - e_k\| \tag{2}$$

The residual for the next level is computed as

$$r_{d+1} := r_d - e_{c_d} \tag{3}$$

This process is recursively repeated m times to generate a tuple of m codewords representing the Semantic ID, approximating the input from coarse to fine granularity. Separate codebooks for each level allow for varying granularities as the residual norms decrease.

3.2 Unified Feature Representation Space

To ensure a consistent and structured representation of items across different domains, we first standard-

Algorithm 1 Training Process of CL-VAE

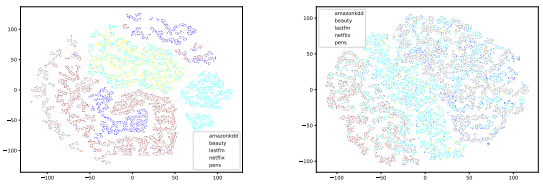
Input: semantic embeddings $e \in \mathbb{R}^{B \times D}$, encoder $\mathcal{E}$, decoder $\mathcal{D}$, m-level codebooks $\{C_d\}_{d=1}^{m}$

used_level=0
Encoding:
$\mu, \log\sigma^2 \leftarrow \mathcal{E}(e)$
$z \leftarrow \mu + \sigma \odot \epsilon$ {Sampling $\epsilon \sim \mathcal{N}(0, I)$}
Curriculum Learning:
if Total loss converges **then**
    used_level $\leftarrow$ used_level $+ 1$
end if
Quantization:
z_q, indices, $L_Q \leftarrow \text{rqvae}(z, \{C_d\}_{d=0}^{\text{used\_level}})$
Codebook Initialization:
if codebook usage rate is low **then**
    $C_0 \leftarrow \text{KMeans}(z)$
end if
Loss Computation:
$L_e \leftarrow \text{entropy\_loss}(\text{indices})$
$e_{\text{recon}} \leftarrow \mathcal{D}(z_q)$
$L_{\text{R}} \leftarrow \text{MSE}(e_{\text{recon}}, e)$
$L_{total} \leftarrow L_{\text{R}} + L_Q + \gamma L_e$

ize item descriptions into a unified format. This allows for uniform processing and facilitates transformation into feature embeddings using a shared encoder. By maintaining structural consistency, the model can effectively learn and compare items across domains within a common feature space.

For a highly generalizable large language recommendation model, it is essential that IDs derived from item embeddings are evenly distributed across the ID space. This ensures that the representations of items from diverse domains are well spread out, making it easier for the model to generalize across various categories. However, during RQ-VAE training, codebook collapse often occurs, where the majority of inputs, regardless of their domain, are mapped to only a small subset of codebook vectors. This results in a situation where a new item, especially one from a previously unseen domain, may be encoded into a token that the large language model (LLM) has never encountered, leading to suboptimal performance. To mitigate this issue, we propose **Curriculum Learning Enhanced RQ-VAE (CL-VAE)**, which improves the model's robustness and enhances its ability to handle zero-shot scenarios across diverse domains.

As shown in Figure 1, in CL-VAE, we introduce curriculum learning (Bengio et al., 2009; Wang et al., 2021b), where the core idea is to allow the model to progressively learn tasks from simple to complex (Narvekar et al., 2020), rather than directly learning complex tasks. We observed that the hierarchical structure of RQ-VAE inherently aligns with the principles of curriculum learning. We perform staged training for the different levels of the codebook: initially, we train the first layer for n epochs to allow it to fully learn basic feature representations, and after the loss stabilizes, we add the second layer for further training, and so on. This staged training approach not only reduces the complexity of initial training but also enhances the model's convergence stability. By gradually building hierarchical representations, CL-VAE effectively maps item representations from diverse domains into a unified concept ID space, enabling the large language model to generalize more effectively across different distributions.



(a) RQ-VAE (b) CL-VAE

Figure 2: Visualization of t-SNE Clustering in the ID Space: (a) Discrete ID space learned by RQ-VAE, (b)Discrete ID space learned by CL-VAE.

To further mitigate codebook collapse, particularly in cases where certain codebook vectors are sparsely utilized, we determine whether to reinitialize the first-level codebook based on its usage during training. This reinitialization provides a new optimization starting point for the first-level codebook, ensuring sufficient learning of low-level features and preventing collapse, thereby enhancing the overall performance of the model.

The loss for the modified model is composed of reconstruction loss $\mathcal{L}_{\text{R}}$, codebook loss and commitment loss $\mathcal{L}_{\text{Q}}$, entropy loss $\mathcal{L}_{\text{E}}$:

$$\mathcal{L}(x) := \mathcal{L}_{\text{R}} + \mathcal{L}_{\text{Q}} + \gamma\mathcal{L}_{\text{E}}, \tag{4}$$

$$\mathcal{L}_{\text{R}} := |x - \hat{x}|^2, \tag{5}$$

$$\mathcal{L}_{\text{Q}} := \sum_{d=0}^{m-1} |\,\text{sg}[r_i] - e_{c_i}|^2 + \beta|r_i - \text{sg}[e_{c_i}]|^2, \tag{6}$$

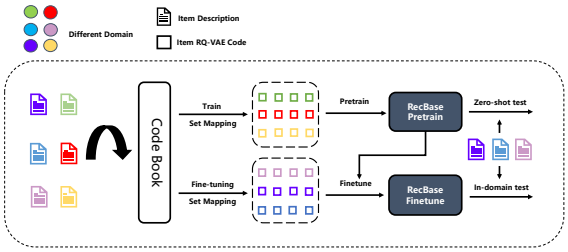


Figure 3: Illustration of the zero-shot transfer and domain-specific fine-tuning process. The diagram shows how the pretrained model is tested in a zero-shot setting and fine-tuned for in-domain performance.

$$\mathcal{L}_{\mathrm{E}} := -\sum_{d=0}^{m-1}\sum_{j=1}^{K} p_{d,j}\log p_{d,j}. \tag{7}$$

here $\hat{x}$ is the output of the decoder, sg denotes the stop-gradient operation, and $p_{d,j}$ represents the usage frequency of codebook vector e_j at level d. The first two loss terms are intrinsic to RQ-VAE, while the third term is an entropy penalty designed to promote more diverse utilization of the codebook. Together, these loss terms facilitate the joint training of the encoder, decoder, and codebook.

3.3 Autoregressive Modeling

After applying the unified discretization method, we obtain a sequence of concept IDs, where each item is represented as an m-bit semantic ID. To utilize this representation for recommendation, we convert a user's interaction history into a sequence of concept IDs, preserving the chronological order. Each item in the sequence is denoted as $s_i = (s_i^1, s_i^2, \ldots, s_i^m)$, where s_i^j is the j-th bit of the item's concept ID.

Given a sequence of historical interactions $S = (s_1, s_2, \ldots, s_n)$, we train an autoregressive model to predict the next ID. The model takes the preceding sequence $S_{<t}$ as input and outputs a probability distribution over each bit of the next ID s_t:

$$P(s_t|S_{<t}) = \prod_{j=1}^{m} P(s_t^j|s_t^{<j}, S_{<t}), \tag{8}$$

where $P(s_t^j|s_t^{<j}, S_{<t})$ is the probability of the j-th bit given its previous bits and the interaction history. Training is performed using the negative log-likelihood loss:

$$\mathcal{L} = -\sum_{t=1}^{n}\sum_{j=1}^{m} \log P(s_t^{j*}|s_t^{<j*}, S_{<t}), \tag{9}$$

where s_t^{j*} is the ground truth bit value.

During inference, the model generates the next item's concept ID bit by bit, treating concept IDs as tokens in its vocabulary. This structured representation enables the model to effectively capture user behavior patterns, leading to more accurate and diverse recommendations.

4 Experiments

We evaluate our method through zero-shot recommendation experiments, pre-training our model on 15 diverse datasets and testing it on 8 unseen datasets to assess its generalization ability. The dataset statistics are shown in Figure 4 and Table 2.

4.1 Evaluation Setup

RecBench (Liu et al., 2025a,b) is a recently introduced benchmark (comprising 15 datasets across 10 domains) designed to evaluate LLM-based recommenders. In this work, we adopt RecBench for zero-shot recommendation evaluation in the majority of our experiments. More specifically, we use item textual information to represent each item, including that in the user browsing history. Given a user–item pair, we will use natural language to concatenate the user and item feature: "The user has browsed the following items: ... , will this user be interested in the item: ...? Answer (Yes/No): ". The prediction of the next token will be regarded as the user interest score or click probability). For closed-source models, we map the LLM's textual responses (YES or NO) directly to interest scores of 1.0 and 0.0, respectively. For open-source models, we obtain the logits of the YES and NO tokens from the classifier, denoted as l_{yes} and l_{no}. After applying softmax normalization **over these two tokens**, we take the score corresponding to the `YES` token as the click probability, formulated as:

$$\text{Click Probability} = p_{\text{yes}} = \frac{e^{l_{\text{yes}}}}{e^{l_{\text{yes}}} + e^{l_{\text{no}}}}. \tag{10}$$

For inference, RecBase takes the user's historical interaction sequence as input and outputs the logits for the predicted item concept IDs. These logits represent the joint probability distribution over the m possible concept IDs. By comparing these probabilities, we identify the item that the user is most likely to engage with, based on their previous interactions. This process allows RecBase to generate recommendations by ranking items according to their predicted interest scores.

Table 1: Zero-shot recommendation evaluation across multi-domain datasets.

	Size(M)	MIND	MovieLens	MicroLens	Goodreads	Yelp	Steam	H&M	HotelRec	Overall
P5	223	0.4911	0.5138	0.5017	0.5027	0.5080	0.5296	0.4845	0.4905	0.5027
RecGPT	6,649	0.5078	0.5069	0.4703	0.5083	0.5140	0.4924	0.4875	0.4937	0.4976
$\text{BERT}_{\text{base}}$	110	0.4963	0.4934	0.4992	0.4958	0.4914	0.5002	0.5204	0.4955	0.4990
OPT_{base}	331	0.5490	0.5104	0.4773	0.5015	0.5158	0.4257	0.4555	0.5028	0.4922
$\text{OPT}_{\text{large}}$	1,316	0.5338	0.5174	0.5236	0.5042	0.5026	0.3825	0.5650	0.5026	0.5039
Qwen-2	494	0.4886	0.5138	0.5701	0.5148	0.5077	0.6399	0.6287	0.5311	0.5493
Phi-2	2,780	0.4851	0.5296	0.5078	0.5049	0.5186	0.6061	0.5447	0.4986	0.5244
Llama-2	6,738	0.4945	0.6030	0.4877	0.5273	**0.5378**	0.5622	0.4519	0.5305	0.5243
Llama-3	8,030	0.4904	0.6412	0.5577	0.5191	0.5267	0.7690	0.5454	0.5342	0.5729
Mistral	7,248	0.4833	**0.6933**	0.559	0.5321	0.5313	0.8102	0.5762	0.5677	0.5941
Deepseek-Qwen2	7,615	0.5117	0.5407	0.563	0.5165	0.5303	0.5905	0.5994	**0.5648**	0.5520
GPT-3.5	-	0.5057	0.5170	0.5110	0.5122	0.5039	0.6184	0.5801	0.5076	0.5319
$\text{RecBase}_{\text{base}}$	313	**0.5508**	0.5352	0.5401	0.5029	0.5320	0.7450	0.5870	0.4874	0.5601
$\text{RecBase}_{\text{large}}$	1,318	0.5442	0.6474	**0.5712**	**0.5329**	0.5326	**0.8343**	**0.6761**	0.5124	**0.6063**

The task is framed as predicting user interest in unseen items based on historical interactions, approached as a ranking problem. To evaluate the model's performance, we use the **Area Under Curve (AUC)** metric, which measures the model's ability to rank items in order of relevance.

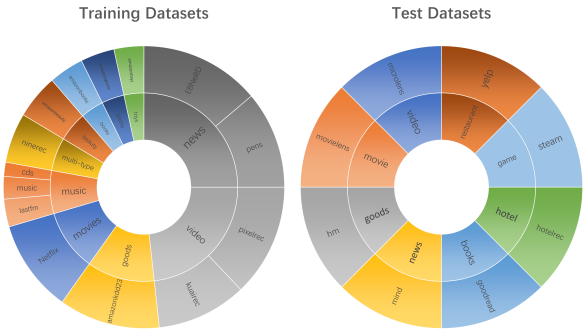


Figure 4: Training and test datasets distribution

Table 2: Statistics of training, finetune, and test datasets.

	Item size	User size	History Avg. length
Training datasets	4,595,003	35,047,682	20.37
Finetune datasets	1,005,745	5,098,084	17.83
Test datasets	623,615	145,975	15.01

4.2 Implementation Details

In our approach, we employ a hierarchical feature extraction technique using the CL-VAE model with a 4-level codebook, each level sized at 2048. This setup enables the model to progressively extract more structured features from raw data, as validated by ablation studies. For textual item descriptions, we use the NV-Embed-v2 (Lee et al., 2024) model to convert unstructured text into dense, semantically rich embeddings, making the data suitable for downstream recommendation tasks.

For the RecBase pre-training, we train two versions: Base and Large. The Base version is configured with a hidden size of 1024, an intermediate size of 2816, and 16 attention heads across 24 layers. It supports a maximum position embedding length of 32,768. The Large version increases the hidden size to 1536, intermediate size to 8960, and adjusts the number of attention heads to 12 with 28 layers. The position embedding length and sliding window are extended to 131,072, enhancing the model's capacity to process longer sequences. Both models use a vocabulary size of 20,000 and share other key settings derived from the Qwen2 (Yang et al., 2024) architecture.

4.3 Zero-Shot Recommendation Performance

In this experiment, we evaluate our RecBase model against several state-of-the-art approaches, including zero-shot recommendation methods based on large language models (LLMs) and fine-tuned recommendation models. Specifically, we compare our model with LLM-based zero-shot methods such as $\text{BERT}_{\text{base}}$ (Devlin et al., 2019), OPT (Zhang et al., 2022), Qwen-2 (Yang et al., 2024), Phi-2, Llama-2 (Touvron et al., 2023), Llama-3 (Dubey et al., 2024), Mistral (Jiang et al., 2023), and GPT-3.5 (Brown et al., 2020), as well as fine-tuned LLM-based models like RecGPT (Zhang et al., 2024) and P5 (Geng et al., 2022). This com-

Table 3: Performance of our model under zero-shot and fine-tuning settings on various datasets.

	Microlens	Steam	MovieLens	H&M	Yelp
Zero-shot	0.5401	0.7450	0.5352	0.5870	0.5320
Fine-tuned	0.5602	0.9173	0.6216	0.6261	0.6125
Improve. (%)	**3.70%**	**23.12%**	**16.14%**	**6.66%**	**15.13%**

parison is conducted across multiple datasets to evaluate the generalization ability of each method.

The experimental results demonstrate the superiority of RecBase over traditional LLM-based methods in both generalization and efficiency. Specifically, RecBaselarge achieves an overall score of 0.6063, surpassing strong baselines such as RecGPT, P5, Mistral, and GPT-3.5. These findings highlight the effectiveness of our domain-specific pretraining strategy, which enables the model to capture fine-grained semantic nuances of recommended items and achieve robust zero-shot performance. Moreover, RecBase consistently outperforms advanced LLM-based models such as Llama-3 and Qwen-2 across all datasets, with particularly notable improvements on H&M (0.6761 vs. 0.6287) and Steam (0.8343 vs. 0.8102), underscoring its strong generalization capability. In addition, RecBasebase, with only 313M parameters, delivers competitive results while outperforming models like BERTbase and OPTbase, all at a substantially lower computational cost. Collectively, these results position RecBase as both a highly efficient and powerful solution for zero-shot recommendation ranking.

4.4 Unified Representation Performance

To evaluate the effectiveness of the unified concept space generated by the CL-VAE method, we conducted an in-depth analysis. As previously noted, an optimal code discretization approach should aim to map the input data distribution as uniformly as possible into the latent space, maximizing the utilization of each token. This ensures that the autoregressive model can effectively leverage every token value for encoding. As illustrated in Figure 2, traditional methods such as RQ-VAE result in features from different domains being independently distributed in the latent space, with minimal interaction between them. In contrast, the CL-VAE method significantly improves upon this, as evidenced by the increased overlap and interaction between features from diverse datasets within the unified concept space.

The transformation in the frequency distribution of code usage across different levels, as depicted in Figure 5c, further supports the validity of this mapping. The CL-VAE method not only achieves a more balanced distribution of IDs but also ensures the effective exploitation of the hierarchical structure of the codebook. This hierarchical approach enables the model to capture both fine-grained and coarse-grained features, thereby enhancing its generalization capability across various recommendation scenarios. By establishing a unified and well-distributed concept space, CL-VAE facilitates the autoregressive model in making more efficient and accurate predictions, ultimately improving the overall performance of the recommendation system.

5 Analysis

In this section, we present a series of ablation studies to evaluate the contribution of individual components to the model's overall performance and generalization ability.

5.1 Ablation Analysis on Key Components

The ablation study, as shown in Table 4, highlights the importance of each component within CL-VAE. The baseline model, $\text{RecBase}_{\text{base}}$, achieves the best performance across all datasets, demonstrating the effectiveness of the complete method. When the formatted text description is removed, there is a noticeable drop in performance, indicating that structured text representations play a vital role in enhancing the model's ability to capture relevant features. Similarly, the removal of the reinitialization step results in a significant decline in performance, suggesting that the initialization mechanism is crucial for stabilizing learning and improving convergence. Additionally, excluding the curriculum learning module leads to further performance degradation, particularly in more complex recommendation scenarios, which underscores the value of progressively training the model on increasingly difficult examples.

5.2 Ablation Study on the Codebook

We perform an ablation analysis on the size and number of levels in the multi-level codebook of the CL-VAE module. As shown in Figure 2, with an increase in size, the conflict rate between the concept IDs obtained from the codebook continuously rises, indicating that enlarging the size benefits the

Table 4: Modular ablation study. format., init. and cur. represent formatted text description, reinitialization and curriculum learning in CL-VAE respectively.

	Yelp	Steam	H&M	HotelRec	Overall
$\text{RecBase}_{\text{base}}$	0.5320	0.7450	0.5870	0.4874	0.5879
w/o format.	0.5204	0.7187	0.5668	0.4966	0.5756
w/o init.	0.4912	0.5924	0.5319	0.4909	0.5266
w/o cur.	0.5073	0.6815	0.5412	0. 4815	0.5529

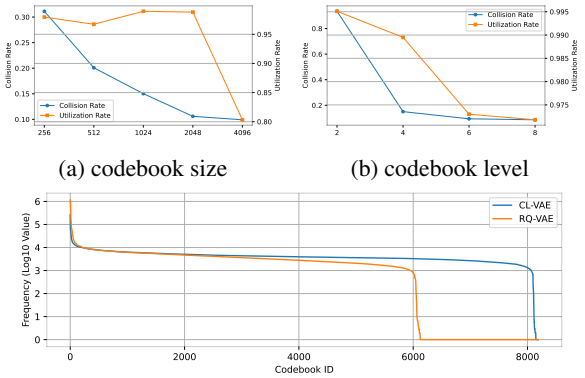
optimization of representations. However, when the size reaches a certain scale, such as 4096 in the figure, a decrease in utilization occurs, leading to redundant wastage of the vocabulary space. Therefore, our model selects a size of 2048 for each layer. Regarding the analysis of levels, as the number of levels increases, the number of products that can be represented by the concept IDs grows exponentially. For spaces beyond four levels, the utilization of IDs becomes very low, and the gains from increasing the number of levels begin to plateau, further incurring additional inference costs during decoding. Consequently, our model adopts a strategy of four levels to optimize the balance between performance and efficiency.

5.3 In-Domain Adaptation via Fine-Tuning

Table 3 shows the performance of our model under zero-shot and fine-tuning settings across various datasets. Fine-tuning consistently improves performance over the zero-shot setting, with significant gains on the Steam (+0.1723) and MovieLens (+0.0864) datasets, indicating the effectiveness of domain-specific adaptation. Even datasets with smaller improvements, such as Microlens (+0.0201) and Yelp (+0.0805), benefit from fine-tuning. Figure 3 illustrates the fine-tuning process, emphasizing how in-domain adaptation refines the model's representations. These results demonstrate the adaptability and potential of our model for further performance enhancement with in-domain supervision.

5.4 Analysis of Inference Efficiency

We conducted an inference efficiency comparison between our model and several other state-of-the-art models, including Qwen2, phi, GptRec, Mistral, and RecBase, on the same experimental dataset. Our model consistently demonstrated superior inference efficiency, outperforming the others by a significant margin. This improvement can be attributed to the specialized ID vocabulary space we designed specifically for recommendation tasks. Unlike general-purpose large language models that rely on vast vocabulary spaces based on natural language representations, our model utilizes a much smaller, more efficient vocabulary tailored to the needs of recommendation systems. This design choice not only enhances the model's efficiency but also positions it as a more suitable base model for recommendation-related tasks.



(a) codebook size (b) codebook level

(c) Codebook usage frequency in CL-VAE and RQ-VAE.

Figure 5: Codebook structure vs. Collision rate and Utilization rate. (a) demonstrates the impact of codebook size on the collision rate and utilization rate. (b) reflects the influence of codebook level on the aforementioned metrics.

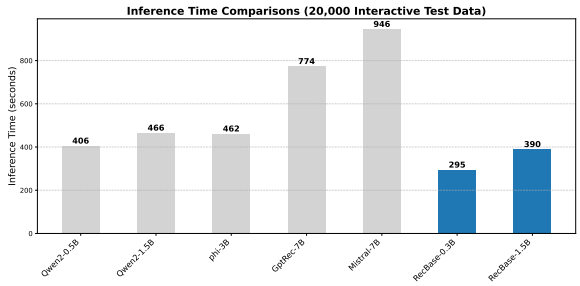


Figure 6: Comparison of Inference Latency.

6 Conclusion

In this paper, we introduce RecBase, a foundation model tailored to the challenges of zero-shot and multi-domain recommendation. By pretraining on a large-scale cross-domain corpus with structured text representations and unified feature mappings, RecBase demonstrates strong generalization across heterogeneous recommendation tasks. The incorporation of curriculum learning and discrete representations facilitates the construction of a unified concept ID space, thereby mitigating semantic dis-

crepancies between domains. Furthermore, the autoregressive training paradigm enables the model to effectively capture inter-item dependencies, yielding superior zero-shot and cross-domain performance compared to conventional large language models. Comprehensive evaluation on eight real-world datasets confirms the effectiveness of our approach, particularly in cold-start scenarios, highlighting the potential of recommendation-oriented pretraining as a promising direction for building robust and adaptable recommender systems.

7 Limitation

Despite its promising performance in zero-shot and multi-domain scenarios, our model exhibits several limitations inherent to recommendation data. Data sparsity and distribution imbalance can impair generalization, particularly for cold-start users and long-tail items. While cross-domain pretraining partially alleviates this, the model may still underrepresent certain domains or items with sparse interactions. Additionally, biases in the training data can limit generalization to new domains or diverse user populations. Future work should explore data augmentation, active learning, and bias mitigation strategies, and evaluate the model on larger, more heterogeneous benchmarks to enhance scalability and real-world robustness.

Acknowledgement

This work was supported by the National Key R&D Program of China (2022ZD0162000) and the National Natural Science Foundation of China under Grants No.62222211 and No.U24A20326 . We also acknowledge partial support from MindSpore (`https://www.mindspore.cn`), a new deep learning computing framework.

References

Diego Antognini and Boi Faltings. 2020. HotelRec: a novel very large-scale hotel recommendation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4917–4923, Marseille, France. European Language Resources Association.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

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 2023, Singapore, Singapore, September 18-22, 2023*, pages 1007–1014. ACM.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu Cheng, Yunzhu Pan, Jiaqi Zhang, Yongxin Ni, Aixin Sun, and Fajie Yuan. 2024. An image dataset for benchmarking recommender systems with raw pixels. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 418–426. SIAM.

Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *CoRR*, abs/2205.08084.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International*

Conference on Information & Knowledge Management, pages 540–550.

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 *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022*, pages 299–315. ACM.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Minjie Hong, Yan Xia, Zehan Wang, Jieming Zhu, Ye Wang, Sihang Cai, Xiaoda Yang, Quanyu Dai, Zhenhua Dong, Zhimeng Zhang, and Zhou Zhao. 2025. EAGER-LLM: enhancing large language models as recommenders through exogenous behavior-semantic integration. In *WWW*, pages 2754–2762.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024a. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian J. McAuley, and Wayne Xin Zhao. 2024b. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval (ECIR)*, volume 14609, pages 364–381.

Chengkai Huang, Tong Yu, Kaige Xie, Shuai Zhang, Lina Yao, and Julian J. McAuley. 2024. Foundation models for recommender systems: A survey and new perspectives. *CoRR*, abs/2402.11143.

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2023. Genrec: Large language model for generative recommendation. *CoRR*, abs/2307.00457.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. Eb-nerd a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024*, RecSys Challenge '24, page 1–11. ACM.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *CoRR*, abs/2405.17428.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. *Preprint*, arXiv:2203.01941.

Xingchen Li, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua. 2020. Hierarchical fashion graph network for personalized outfit recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 159–168. ACM.

Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S. Yu, and Kai Shu. 2025. Taxonomy-guided zero-shot recommendations with llms. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 1520–1530.

Jianghao Lin, Bo Chen, Hangyu Wang, Yunjia Xi, Yanru Qu, Xinyi Dai, Kangning Zhang, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Clickprompt: CTR models are strong prompt generators for adapting language models to CTR prediction. In *Proceedings of the ACM on Web Conference 2024 (WWW)*, pages 3319–3330.

Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? A preliminary study. *CoRR*, abs/2304.10149.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024a. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *WSDM*.

Qijiong Liu, Jieming Zhu, Lu Fan, Kun Wang, Hengchang Hu, Wei Guo, Yong Liu, and Xiao-Ming Wu. 2025a. Benchmarking llms in recommendation tasks: A comparative evaluation with conventional recommenders. *arXiv preprint arXiv:2503.05493*.

Qijiong Liu, Jieming Zhu, Yingxin Lai, Xiaoyu Dong, Lu Fan, Zhipeng Bian, Zhenhua Dong, and Xiao-Ming Wu. 2025b. Evaluating recabilities of foundation models: A multi-domain, multi-dataset benchmark. *arXiv preprint arXiv:2508.21354*.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024b. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 6566–6576.

Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. Crosscbr: Cross-view contrastive learning for bundle recommendation. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 1233–1241. ACM.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50.

Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. 2023. A content-driven micro-video recommendation dataset at scale. *arXiv preprint arXiv:2309.15379*.

Bo Peng, Srinivasan Parthasarathy, and Xia Ning. 2022. Recursive attentive methods with reused item representations for sequential recommendation. *CoRR*, abs/2209.07997.

Aleksandr V. Petrov and Craig Macdonald. 2023. Generative sequential recommendation with gptrec. *ARXIV-CS.IR*.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Mahesh Sathiamoorthy. 2023. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning. *Preprint*, arXiv:1711.00937.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610, Florence, Italy. Association for Computational Linguistics.

Shoujin Wang, Longbing Cao, Liang Hu, Shlomo Berkovsky, Xiaoshui Huang, Lin Xiao, and Wenpeng Lu. 2021a. Hierarchical attentive transaction embedding with intra- and inter-transaction dependencies for next-item recommendation. *IEEE Intell. Syst.*, 36(4):56–64.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021b. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and Zhenhua Dong. 2024a. EAGER: two-stream generative recommender with behavior-semantic collaboration. In *KDD*, pages 3245–3254.

Yidan Wang, Zhaochun Ren, Weiwei Sun, Jiyuan Yang, Zhixiang Liang, Xin Chen, Ruobing Xie, Su Yan, Xu Zhang, Pengjie Ren, Zhumin Chen, and Xin Xin. 2024b. Content-based collaborative generation for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 2420–2430. ACM.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

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 (WWW)*, 27(5):60.

An Yang, Baosong Yang, Binyuan Hui, and et al. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Yabin Zhang, Wenhui Yu, Erhan Zhang, Xu Chen, Lantao Hu, Peng Jiang, and Kun Gai. 2024. Recgpt: Generative personalized prompts for sequential recommendation via chatgpt training paradigm. *CoRR*, abs/2404.08675.

Zizhuo Zhang and Bang Wang. 2023. Prompt learning for news recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 227–237. ACM.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*, pages 1435–1448. IEEE.

Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. 2023. Generative job recommendations with large language model. *ARXIV-CS.IR*.

Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. BARS: towards open benchmarking for recommender systems. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2912–2923.

Jieming Zhu, Mengqun Jin, Qijiong Liu, Zexuan Qiu, Zhenhua Dong, and Xiu Li. 2024a. Cost: Contrastive quantization based semantic tokenization for generative recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys)*, pages 969–974.

Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024b. Collaborative large language model for recommender systems. *WWW*.

Supplementary Material

A Dateset Details.

The rec-base model are pretrained on 15 diverse training datasets across various domains and evaluated on 8 additional cross-domain datasets. Here are detail descriptions about these datasets.

Training datasets EBNeRD (Kruse et al., 2024) and PENS (Ao et al., 2021) are news-related datasets. EBNeRD is used to optimize news recommendation systems, while PENS is a dataset for personalized news headline generation. The PixelRec (Cheng et al., 2024) and KuaiRec (Gao et al., 2022) datasets are both related to short video recommendations, containing a large number of user interactions with short videos and the corresponding video thumbnails. The Amazon Reviews 2023 (Hou et al., 2024a) dataset is a large-scale e-commerce review dataset that covers user feedback information such as ratings, review texts, and helpful votes, as well as product metadata. The Amazon Review Dataset records user evaluations of products on the Amazon website and is a classic dataset for recommendation systems. Sub-datasets such as Amazonbeauty, Amazonbooks, Amazonsports, and Amazontoys focus on the fields of beauty products, books, sports products, and toy products, respectively. Netflix dataset is a classic movie-related recommendation system dataset, containing over 100 million user ratings for movies.

Evaluation datasets MIND (Wu et al., 2020) dataset is a large-scale news recommendation dataset constructed from Microsoft News user click logs, featuring rich information for each news article, including title, abstract, body, and category labels. It records users' news click history and impression logs. The dataset is well-suited for studying challenges in news recommendation such as cold start problems, and user interest modeling. MovieLens (Harper and Konstan, 2015) dataset is a classic recommendation system dataset created by the GroupLens research team at the University of Minnesota, containing a large number of movie ratings by users. MicroLens (Ni et al., 2023) is a large-scale, content-driven short video recommendation dataset that contains 1 billion user interactions with short videos and provides rich modality information for these videos. The Goodreads (Wan et al., 2019) dataset was collected from the Goodreads website by UC San Diego, containing book metadata, user-book interactions, and detailed user reviews, with approximately 228 million user-book interaction records. The Yelp Open Dataset is a subset of business, review, and user data provided by the well-known American merchant review website Yelp. It includes approximately 160,000 businesses, 8.63 million reviews, and 200,000 images from eight major metropolitan areas. The Steam Dataset is a multi-dimensional dataset built based on Steam, the world's largest digital game distribution platform. It covers detailed information such as game purchase records and playtime for millions of users and is widely used for research in user behavior analysis, market trend prediction, and game recommendation systems. The H&M dataset is a dataset provided by H&M, containing product information, customer information, and transaction records. It is widely used for research in recommendation systems. HotelRec (Antognini and Faltings, 2020) is a large-scale hotel recommendation dataset created by the Artificial Intelligence Laboratory at École Polytechnique Fédérale de Lausanne (EPFL), collected from the TripAdvisor platform. It contains approximately 50 million hotel reviews and is the largest recommendation dataset in a single domain with text reviews.

B Data Processing Details

B.1 Biases in Data Distribution and Modality Handling

The RecBase pretraining corpus is constructed from 15 cross-domain recommendation datasets. We acknowledge certain biases in the overall distri-

bution, particularly the overrepresentation of news and audio-visual content. To mitigate this, we carefully curated the dataset selection to achieve a more balanced category distribution, ensuring that no single domain dominates the training signal.

To ensure a controlled and fair evaluation, we standardized all inputs to a text-only format. For multimodal datasets such as PixelRec, Clothing, and NineRec, we therefore isolated and used only their textual descriptions (e.g., product metadata and reviews), while discarding other modalities. This design choice mitigates confounding effects from heterogeneous data types and establishes a level playing field for comparison with our text-centric baselines. While this precludes the use of multimodal signals for now, we argue the rich textual information present in these datasets provides a robust basis for assessing model generalization. The integration of discretized multimodal features into our unified framework is a promising direction for future research.

B.2 Data Preprocessing and Noise Filtering

To maximize exposure to real-world user behavior patterns, RecBase is pretrained on raw user interaction data from each dataset. The preprocessing pipeline includes the following steps:

Text Standardization. We structured item-related content into a unified format, combining titles, attributes, and reviews into a clean textual description. Non-informative or off-topic reviews were filtered out to retain only content relevant to the core product or item, as illustrated in Figure 1(b) and detailed in Appendix B.

User History Filtering. We applied length-based filtering to manage the variance in user history lengths. Users with fewer than 15 interactions (e.g., in *AmazonToys*) were removed due to insufficient sequence signal, while extremely long histories (e.g., exceeding 2500 interactions in *LastFM* or *PixelRec*) were truncated to avoid overfitting and memory inefficiency.

B.3 Negative Sampling

Unlike traditional recommendation models that require explicit negative sampling, RecBase is trained on real user interaction sequences using an autoregressive objective. Therefore, no artificial negative sampling is involved during training. The model learns to predict the next likely item based on historical sequences, leveraging only positive (i.e., observed) user feedback. This setup reflects the natural sequential structure of user behavior and aligns with the pretraining objective.

C Examples of Items.

In what follows, we show some examples of formatted text descriptions of items for extracting embeddings by using the NV-Embedding model (Lee et al., 2024):

'Describe a movie:\n{\n"title": "Toy Story (1995)",\n"genres": "Adventure | Animation | Children | Comedy | Fantasy"\n}'

'Describe a movie:\n{\n"title": "Jumanji (1995)",\n"genres": "Adventure | Children | Fantasy"\n}'

'Describe a movie:\n{\n"title": "Grumpier Old Men (1995)",\n"genres": "Comedy | Romance"\n}'

Test cases for large language model benchmarks:

'User behavior sequence: \n(1) This Ford GT40 Movie Rig From "Ford V Ferrari" Looks Absurd\n(2) Kendall Jenner Wore the Tiniest Dress to Go Jewelry Shopping\nCandidate item: 9 fashion trends inspired by the 2000s that are coming back in style'

'User behavior sequence: \n(1) This Ford GT40 Movie Rig From "Ford V Ferrari" Looks Absurd\n(2) Kendall Jenner Wore the Tiniest Dress to Go Jewelry Shopping\nCandidate item: Here Are the Biggest Deals Weŕe Anticipating for Black Friday'

'User behavior sequence: \n(1) This Ford GT40 Movie Rig From "Ford V Ferrari" Looks Absurd\n(2) Kendall Jenner Wore the Tiniest Dress to Go Jewelry Shopping\nCandidate item: Man cuffed for eating sandwich on train platform gets an apology'