

MiniOneRec: An Open-Source Framework for Scaling Generative Recommendation

Xiaoyu Kong*, Leheng Sheng^o, Junfei Tan*, Yuxin Chen^o,
Jiancan Wu*, An Zhang*, Xiang Wang*, Xiangnan He*

*University of Science and Technology of China ^oNational University of Singapore

{kongxy, sober_clever}@mail.ustc.edu.com,

{leheng.sheng, yuxin.chen}@u.nus.edu,

{wujcan, an.zhang3.14, xiangwang1223, xiangnanhe}@gmail.com

ABSTRACT

The recent success of large language models (LLMs) has renewed interest in whether recommender systems can achieve similar scaling benefits. Conventional recommenders, dominated by massive embedding tables, tend to plateau as embedding dimensions grow. In contrast, the emerging generative paradigm replaces embeddings with compact Semantic ID (SID) sequences produced by autoregressive Transformers. Yet most industrial deployments remain proprietary, leaving two fundamental questions open: (1) Do the expected scaling laws hold on public benchmarks? (2) What is the minimal post-training recipe that enables competitive performance?

We present **MiniOneRec**, to the best of our knowledge, the first fully open-source generative recommendation framework, which provides an end-to-end workflow spanning SID construction, supervised fine-tuning, and recommendation-oriented reinforcement learning. We generate SIDs via a Residual Quantized VAE and post-train Qwen backbones ranging from 0.5B to 7B parameters on the Amazon Review dataset. Our experiments reveal a consistent downward trend in both training and evaluation losses with increasing model size, validating the parameter efficiency of the generative approach. To further enhance performance, we propose a lightweight yet effective post-training pipeline that (1) enforces full-process SID alignment and (2) applies reinforcement learning with constrained decoding and hybrid rewards. Together, these techniques yield significant improvements in both ranking accuracy and candidate diversity.

🔗 <https://github.com/AkaliKong/MiniOneRec>

🗨️ <https://huggingface.co/kkknight/MiniOneRec>

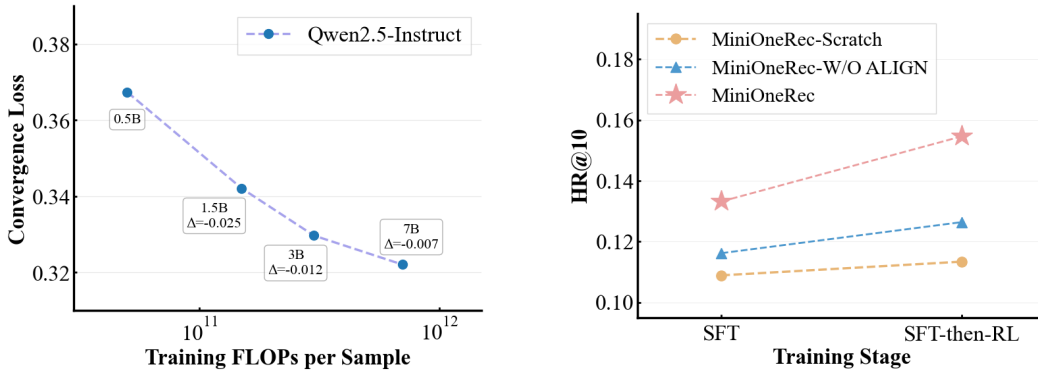


Figure 1: Left: Scaling curves from 0.5B to 7B parameters. Right: Effect of world knowledge on model performance: MiniOneRec-W/O ALIGN uses pretrained LLM weights but omits SID-text alignment, while MiniOneRec-Scratch is trained from random initialization and omits alignment.

CONTENTS

1	Introduction	3
2	Background and Related Work	4
2.1	Generative Recommendation	4
2.2	LLM and RL	4
3	Modeling	5
3.1	Task formulation	5
3.2	Item Tokenization	5
3.3	Alignment with LLMs	6
3.4	Reinforced Preference Optimization	6
3.4.1	Sampling Strategy	7
3.4.2	Reward Design	7
4	Training	8
5	Evaluation	8
5.1	Evaluation Setup	9
5.2	Scaling	9
5.3	Performance Comparison	9
5.4	Transferability	10
5.5	Ablation Study	10
5.5.1	Aligning Strategy	10
5.5.2	Sampling Strategy	11
5.5.3	Reward Design	11
5.6	Pre-trained LLM Impact	12
6	Conclusion	12
A	Alignment Prompts	18
A.1	Overview	18
A.2	Recommendation Tasks	18
A.3	Alignment Tasks	18
B	Datasets	19

1 INTRODUCTION

The scaling behavior of recommendation models has recently attracted significant attention (Deng et al., 2025; Wang et al., 2025; Zhai et al., 2024; Zhang et al., 2024; Jiang et al., 2025; Zhu et al., 2025; Chen et al., 2024a; Dai et al., 2025), driven largely by the success of large language models (LLMs) in demonstrating predictable performance gains with increased model size (Brown et al., 2020; Achiam et al., 2023; DeepSeek-AI et al., 2024; Yang et al., 2024; Jiang et al., 2023; Dubey et al., 2024; Touvron et al., 2023; DeepSeek-AI et al., 2025). However, traditional recommendation models have struggled to exhibit similar scaling laws (Kang and McAuley, 2018; Hidasi et al., 2016; Tang and Wang, 2018; Wu et al., 2021; Fang et al., 2020). These systems typically allocate the majority of parameters to large embedding tables for storing user and item representations, while using only inner-product or shallow scoring networks for final predictions. Such an embedding-heavy design leads to performance plateaus: even as embedding dimensions or table sizes increase, improvements diminish quickly beyond moderate scales (Zhou et al., 2018; Covington et al., 2016; Wang et al., 2021; Guo et al., 2017).

Generative recommendation offers a fundamental paradigm shift. By compressing items into sequences of discrete Semantic IDs (SIDs) through quantization techniques (Luo et al., 2024; Zeghidour et al., 2022), it uses compact vocabularies and redirects the bulk of parameters toward deep autoregressive Transformers that generate SID sequences (Rajput et al., 2023; Zheng et al., 2024; Qu et al., 2024; Deng et al., 2025; Dai et al., 2025). This design enables scaling behaviors more characteristic of language models, where increased depth and capacity translate to consistent performance gains (Deng et al., 2025).

Recent industrial deployments have demonstrated the promise of this generative paradigm. OneRec (Deng et al., 2025; Zhou et al., 2025a) achieves significant improvements on Kuaishou’s platform with 400 million daily active users, while OneRec-V2 (Zhou et al., 2025b) further advances through lazy decoder architecture and preference alignment. OnePiece (Dai et al., 2025) integrates LLM-style context engineering and reasoning into both retrieval and ranking models of industrial cascaded pipelines. However, these results rely on massive proprietary datasets and remain closed-source, leaving critical questions unanswered for the research community:

- Do the scaling advantages of generative recommendation transfer to public datasets?
- What is the minimal post-training recipe needed to achieve strong performance?

In this work, we present **MiniOneRec**, the first fully open-source framework for generative recommendation, which provides an end-to-end workflow encompassing SID generation, model supervised fine-tuning (SFT), and recommendation-driven reinforcement learning (RL). The release includes complete source code, reproducible training pipelines, and publicly available model checkpoints. We implement SID construction using residual quantized variational autoencoder (RQ-VAE) (Zeghidour et al., 2022), in line with prior work such as TIGER (Rajput et al., 2023; Deng et al., 2025). We post-train Qwen-based (backbone) generative models at multiple scales, ranging from 0.5B to 7B parameters, on public benchmarks from the Amazon Review Dataset. Following OneRec’s successful two-stage post-training pipeline, we adopt SFT on user-item interaction sequences, followed by group relative policy gradient (GRPO) (DeepSeek-AI et al., 2025; Shao et al., 2024) for alignment. The main contributions of our work are summarized as follows:

- **Validating the Scaling Laws of Generative Recommendation:** We present the first systematic investigation of how generative recommendation models scale on public data, to the best of our knowledge. Specifically, we train four MiniOneRec variants, spanning 0.5 B to 7 B parameters, on the Amazon Review corpus (Hou et al., 2024) with identical data splits and hyper-parameter settings. Both the final training loss and the held-out evaluation loss consistently decrease as model size grows (see Figures 1 Left and 3), highlighting the superior parameter efficiency of the generative paradigm.
- **Optimizing Post-training Strategies for Generative Recommendation:** We design a lightweight yet comprehensive post-training pipeline that tightly couples full-process SID alignment with reinforced preference optimization. First, we verify the influence of world knowledge on the model’s generative recommendation performance (see Figure 1 Right) and augment the vocabulary with dedicated SID tokens and enforce auxiliary alignment objectives throughout the two-stage optimization process. Second, during the RL phase,

we shape the generation process itself: invalid tokens are masked to guarantee that every step produces a legal item, beam search is employed for efficient exploration of diverse candidates, and the reward signal combines rule-based accuracy with a ranking-aware penalty that pushes the model away from hard negatives. As a result, MiniOneRec offers a concise yet effective recipe for bringing RL with value-based ranking into generative recommendation, simultaneously improving candidate diversity and ranking fidelity.

The remainder of this report is organized as follows: Section 2 reviews background on generative recommendation and the application of LLMs with RL in recommender systems. Section 3 presents our proposed MiniOneRec, including task formulation, item tokenization, and post-training pipeline. Section 4 describes the training details while Section 5 provides comprehensive experimental results. Section 6 concludes with discussions of future directions.

2 BACKGROUND AND RELATED WORK

2.1 GENERATIVE RECOMMENDATION

In the last few years, formulating recommendation as a sequence generation problem has become a vibrant research direction (Rajput et al., 2023; Hou et al., 2025). Under this view, a recommender is trained to predict the tokens of the next item in an end-to-end manner, typically with a Transformer backbone (Vaswani et al., 2017). Such a design removes the rigid multi-stage “matching–ranking” pipeline in traditional retrieval-based recommender systems (Kang and McAuley, 2018; Hidasi et al., 2016) and therefore pushes the performance upper bound.

Early explorations illustrate two key components: (1) turning an item into a discrete code (SIDs) and (2) letting the model produce the code. TIGER (Rajput et al., 2023) adopts RQ-VAE (Zeghidour et al., 2022) to map textual embeddings of titles and descriptions into SIDs. HSTU (Zhai et al., 2024) introduces a streaming architecture that is friendly to high-cardinality and non-stationary logs. LC-Rec (Zheng et al., 2024) aligns an LLM with the SIDs through multi-task learning so that the model can understand these symbols during generation.

Subsequent work focuses on designing better codes. RecForest (Feng et al., 2022) clusters items via hierarchical k -means and uses the cluster indices as tokens. EAGER (Wang et al., 2024) and TokenRec (Qu et al., 2024) fuse collaborative and semantic evidence directly into the tokenizer.

At industrial scale, generative recommenders have started to replace heavy cascade systems. MTGR (Wang et al., 2025) keeps the original DLRM features, adds user-level compression, and accelerates both training and inference. OneRec (Zhou et al., 2025b) reduces serving cost through a lazy decoder-only layout and stabilizes optimization with an improved RL algorithm. OnePiece (Dai et al., 2025) discovers that performing inference in the latent space can further improve generative recommendation performance.

2.2 LLM AND RL

RL aims at maximizing cumulative reward through repeated interaction (Kaelbling et al., 1996). When fine-tuning LLMs, RL with Human Feedback (RLHF) has become a standard recipe; PPO (Schulman et al., 2017) is the most common optimizer but is memory-intensive for billion-scale parameters. To reduce cost, Direct Preference Optimization (DPO) (Rafailov et al., 2023) removes the separate value network and maximizes the log-likelihood gap between preferred and dispreferred outputs. S-DPO (Chen et al., 2024b) adapts this idea to recommendation by treating softmax-based negative sampling as implicit pairwise preference. Nevertheless, preference-based objectives are off-policy and may converge prematurely.

Light-weight on-line methods have therefore been proposed. GRPO (Shao et al., 2024) normalizes rewards within a small batch of roll-outs and replaces the learned reward model by rule-based signals, achieving strong gains in maths and code generation.

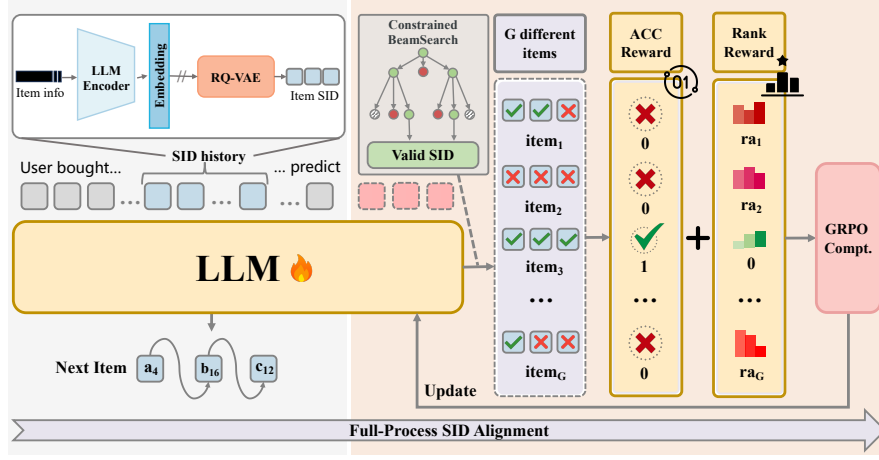


Figure 2: MiniOneRec framework. RQ-VAE builds the item SID codebook. We then perform SFT to warm up the LLM and obtain an initial alignment. In RL, beam search with constrained decoding, thereby the model sequentially produces a ranked list of distinct, valid SIDs. GRPO updates the policy, and SID alignment is enforced end-to-end. This alignment objective is preserved throughout both the SFT and RL stages, fostering deeper semantic understanding.

3 MODELING

In this section, we present the modeling strategies of **MiniOneRec**, as illustrated in Figure 2. MiniOneRec first converts the textual information into SIDs with the RQ-VAE tokenizer (Section 3.2). For better incorporating the huge world knowledge within LLMs (Liao et al., 2023), MiniOneRec further introduces the alignment with LLMs as one expansion of the original OneRec architecture (Section 3.3). During the training stage, MiniOneRec is optimized with next token prediction first, and then followed by reinforced preference optimization (Section 3.4).

3.1 TASK FORMULATION

We first formulate recommendation as a sequence-generation problem. For every user u , the items he or she has interacted with are sorted chronologically to form a sequence $H_u = [i_1, i_2, \dots, i_T]$. Each item i_t in the sequence is encoded by a three-level structural ID, $\{c_0^{i_t}, c_1^{i_t}, c_2^{i_t}\}$. Such structural IDs are typically called SIDs, which preserve hierarchical semantics through quantization techniques with semantic embeddings (Rajput et al., 2023).

A generative policy π_θ , implemented as an autoregressive model with parameters θ , reads the entire history H_u and is trained to predict the next item i^+ that best matches the taste of user u among all candidates in the catalog. During inference, the model recursively produces item tokens; we keep the k most promising beams by the standard beam-search algorithm and return them as the recommendation list. Model performance is reported with the evaluation measures commonly adopted in generative recommendation.

3.2 ITEM TOKENIZATION

In SID-style generative recommenders, the first task is to convert each item into a sequence of discrete tokens. Following the practices of TIGER (Rajput et al., 2023), we employ RQ-VAE (Zeghidour et al., 2022) for this purpose. Concretely, the pipeline is:

1. For every item i , we concatenate its title and textual description to form a single sentence;
2. This sentence is passed through a frozen text encoder (Sheng et al., 2025), producing a d -dimensional semantic vector $\mathbf{x} \in \mathbb{R}^d$;
3. Apply RQ-VAE to \mathbf{x} . At each level l ($0 \leq l < L$) we have a separate codebook $\mathcal{C}_l = \{\mathbf{e}_k^{(l)}\}_{k=1}^K$, where K is the codebook size. We set $L=3$ and $K=256$, so each item is

represented by three bytes; this choice provides 2^{24} possible codes, which is sufficient for catalogs containing hundreds of millions of products while keeping the vocabulary small. The residual is initialized as $\mathbf{r}_0 = \mathbf{x}$ and updated by

$$c_l = \arg \min_k \left\| \mathbf{r}_l - \mathbf{e}_k^{(l)} \right\|_2, \quad \mathbf{r}_{l+1} = \mathbf{r}_l - \mathbf{e}_{c_l}^{(l)}.$$

4. Collect the indices (c_0, \dots, c_{L-1}) as the discrete token sequence for item i ; these indices constitute the item tokens consumed by the subsequent generative recommender.

The quantized latent is reconstructed by

$$\mathbf{z}_q = \sum_{l=0}^{L-1} \mathbf{e}_{c_l}^{(l)}, \quad \hat{\mathbf{x}} = D(\mathbf{z}_q),$$

where $D(\cdot)$ is a decoder. The codebooks, the encoder and the decoder are trained jointly, while the text encoder remains frozen. The loss is the sum of a reconstruction term and an RQ regularizer:

$$\mathcal{L}(\mathbf{x}) = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\mathcal{L}_{\text{RECO}}} + \underbrace{\sum_{l=0}^{L-1} (\|\text{sg}[\mathbf{r}_l] - \mathbf{e}_{c_l}^{(l)}\|_2^2 + \beta \|\mathbf{r}_l - \text{sg}[\mathbf{e}_{c_l}^{(l)}]\|_2^2)}_{\mathcal{L}_{\text{RQ}}},$$

with $\text{sg}[\cdot]$ the stop-gradient operator and β a hyper-parameter controlling the commitment term.

To prevent codebook collapse, we follow the warm-start trick in prior work (Zeghidour et al., 2022; Zheng et al., 2024) and initialize each codebook with k -means centroids computed on the first training batch.

3.3 ALIGNMENT WITH LLMs

LLMs possess extensive understanding of the world and human behaviors (Sheng et al., 2025), which can serve as a supplement to the collaborative signals in recommender systems (Ren et al., 2024). Existing work has shown that linking an LLM’s world knowledge to SID representations noticeably strengthens generative recommendation (Liao et al., 2023; Zheng et al., 2024). Therefore, rather than training on SIDs alone as in prior work (Rajput et al., 2023; Deng et al., 2025; Wang et al., 2025), we introduce several alignment objectives that tie the language space to SID signals. Two major groups of tasks are employed:

- **Recommendation Tasks:** The LLM receives a time-ordered history together with a clear instruction and is asked to predict the SID of the next item the user might engage with.
- **Alignment Tasks:** A collection of bridging tasks enforces a two-way mapping between natural language and SID space, grounding the discrete codes in text while injecting linguistic knowledge into their embeddings.

Tasks from both groups are optimized jointly throughout the SFT stage and the subsequent RL stage. During RL, we adopt constrained decoding so that the model can only produce tokens from a predefined list containing every item’s SID and its canonical title. This constraint guarantees valid outputs and enables straightforward, rule-based reward computation. Detailed examples of the prompts are provided in the Appendix A.

3.4 REINFORCED PREFERENCE OPTIMIZATION

After SFT, we further polish the policy with GRPO (Shao et al., 2024; DeepSeek-AI et al., 2024). GRPO differs from classic RLHF in that it draws multiple candidates per prompt and normalizes rewards within the group, which reduces gradient variance.

Steps: (1) For every prompt $x \sim D$, the frozen policy $\pi_{\theta_{\text{old}}}$ is rolled out G times, yielding $\mathcal{Y}(x) = \{y^{(1)}, \dots, y^{(G)}\}$. (2) Each candidate $y^{(i)}$ is assigned a scalar score S_i . (3) Advantages are standardized inside the group:

$$\hat{A}_i = \frac{S_i - \mu_{1:G}}{\sigma_{1:G}},$$

where $\mu_{1:G}$ and $\sigma_{1:G}$ denote the mean and standard deviation of the G rewards.

The surrogate objective becomes

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, y^{(i)}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y^{(i)}|} \sum_{t=1}^{|y^{(i)}|} \left(\min(w_{i,t} \hat{A}_{i,t}, \text{clip}(w_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t}) - \beta \text{KL}[\pi_\theta \| \pi_{\text{ref}}] \right) \right],$$

with token-level importance ratio $w_{i,t} = \frac{\pi_\theta(y_t^{(i)} | x, y_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)} | x, y_{<t}^{(i)})}$. The parameter ϵ controls clipping, while β keeps the updated policy near a reference model through a KL term.

Applying RL with verifiable rewards (RLVR) to recommendation brings two obstacles:

- **Unique generation space.** The action space is a closed set of item SIDs, orders of magnitude smaller than natural-language vocabularies. Re-sampling tends to produce duplicates, wasting computation. We therefore mix dynamic sampling (Yu et al., 2025) with constrained beam search to enlarge coverage while keeping outputs valid.
- **Sparse ranking supervision.** A hard binary reward (1 for the correct item, 0 otherwise) offers little guidance on ranking quality. We introduce an auxiliary ranking-shaped reward to penalize harder negatives with lower scores. Additionally, dense signals such as semantic similarity and collaborative scores are explored to furnish richer supervision.

3.4.1 SAMPLING STRATEGY

A practical obstacle when porting RLVR to recommendation is the poor sampling diversity brought by the limited action space: querying the policy multiple times with the same prompt often returns identical items, so the model observes few distinct negatives. We measure diversity via

$$\text{Div}(\{e_k\}_{k=1}^G) = \frac{|\text{Unique}(\{e_k\}_1^G)|}{G},$$

where the numerator counts unique items among the G generations. Higher values indicate richer supervision.

Two complementary remedies are investigated:

- **Dynamic Sampling** (Yu et al., 2025). We first over-sample, then pick a subset that (i) must include the ground-truth item and (ii) maximizes internal diversity. Although helpful, this demands extra forward passes and still deteriorates as training progresses.
- **Beam Search.** We ultimately switch to beam search without length normalization (Bao et al., 2024; Tan et al., 2025). By construction, all beams differ, so the method guarantees zero duplication within each group and yields better diversity–efficiency trade-offs.

Based on our findings, MiniOneRec ultimately employs constrained beam search as its default sampler, ensuring that every generated item is valid while still providing a diverse set of candidate trajectories.

3.4.2 REWARD DESIGN

Recommendation models are usually judged with ranking measures such as NDCG. In contrast, the standard GRPO setup supplies a binary reward, giving a value of 1 to the true item and 0 to every other candidate. This strategy treats all negatives as equally harmful. Earlier studies (Wu et al., 2021; Chen et al., 2024b) have shown that focusing on hard negatives produces stronger rankers. Motivated by these results, we introduce a rank-aware reward that assigns different penalties based on how prominently a negative item appears in the model’s own ranking (Tan et al., 2025).

Given a negative candidate e_k whose generation probability ranks ρ_k (with $\rho = 1$ the most probable), we set

$$\tilde{R}_{\text{rank}}(e_k, e_t) = \begin{cases} 0, & e_k = e_t, \\ -\frac{1}{\log(\rho_k + 1)}, & \text{otherwise,} \end{cases} \quad R_{\text{rank}}(e_k, e_t) = -\frac{\tilde{R}_{\text{rank}}(e_k, e_t)}{\sum_{j=1}^G \tilde{R}_{\text{rank}}(e_j, e_t)}.$$

Thus, negatives that the model was very confident about (low ρ_k) receive stronger penalties.

The final reward combines rule-based and ranking components:

$$R(e_k, e_t) = R_{\text{rule}}(e_k, e_t) + R_{\text{rank}}(e_k, e_t),$$

where the rule-based term is

$$R_{\text{rule}}(e_k, e_t) = \begin{cases} 1, & e_k = e_t, \\ 0, & \text{otherwise.} \end{cases}$$

The mixed reward discussed combines binary correctness with a soft ranking term, helping the LLM tell hard negatives apart. Yet recommendation data contain additional, under-utilized cues. To see whether such information can enhance or even replace the rule-based component in RLVR, we experiment with the following collaborative reward. Specifically, for every item suggested by the policy, we obtain its logit from a pre-trained collaborative-filtering model and pass that score back as reward, thereby injecting knowledge extracted from historical user-item interactions. MiniOneRec ultimately adopts the combined ranking-and-rule reward as its default choice.

4 TRAINING

MiniOneRec first converts item text into discrete codes. Titles and descriptions are embedded by the Qwen3-Embedding-4B encoder, after which RQ-VAE performs residual quantization. This tokenizer is trained on a single GPU with a batch size of 20480, a learning rate of 1×10^{-3} , and 10,000 training epochs. The resulting SID vocabulary is plugged into a Qwen2.5-Instruct backbone. SFT takes place on eight NVIDIA H100, each holding 128 samples. Training lasts up to ten epochs with early stopping (patience one epoch). The initial learning rate is 3×10^{-4} and follows a cosine decay schedule. Starting from the SFT checkpoint, we apply GRPO for two additional epochs while keeping the KL weight β unchanged. During roll-out, beam search with width 16 is adopted, so every input generates sixteen distinct candidate sequences.

In the performance comparison with existing recommenders, conventional recommenders are trained with binary cross-entropy and the Adam optimizer; learning rates are chosen from $\{1e-2, 1e-3, 1e-4\}$, and the weight-decay term is scanned over $\{1e-2, 1e-3, 1e-4, 1e-5, 1e-6\}$. The batch size is 1024. For TIGER, we rely on a T5 encoder-decoder and reuse Qwen3-Embedding-4B for item embeddings. All LLM-powered systems, including ours, share the Qwen2.5-Instruct backbone and use AdamW. Batches contain 128 examples for SFT and preference alignment, and 512 samples for RL. The learning rate is 3×10^{-4} during SFT, while S-DPO and RL use 1×10^{-5} . S-DPO runs for one epoch with $\beta = 0.1$ and samples three negative items; D³ tries interpolation factors α of 0.8, 0.9, and 1.0.

5 EVALUATION

This section details our empirical study. We begin with a scaling investigation on two real-world benchmarks (Hou et al., 2024), examining how MiniOneRec’s loss curves change as model size grows. We then benchmark MiniOneRec against three groups of baselines: traditional sequential recommenders, SID-based generators, and recent LLM-driven systems. To evaluate the model’s capability for cross-domain recommendation generalization, we frame SID next-item prediction as a task of recommendation rule discovery and evaluate the model on domains that never appeared during training. Comprehensive ablation tests follow, which help us locate the parts of the framework that contribute most to the final score. Finally, we look into how the broad knowledge stored in pre-trained LLM weights affects generative recommendation. In summary, we study the following questions within this section:

- How does the loss of MiniOneRec scale with different model size?
- How does MiniOneRec perform in comparison to other baseline methods?
- How does MiniOneRec perform under completely unseen domains?
- How do the designed components contribute to MiniOneRec’s recommendation efficiency?
- How do the pre-trained weights of the LLM influence the performance of generative recommendation?

Table 1: Performance of MiniOneRec Compared to Traditional Methods, Generative Methods, and LLM-based Methods

Datasets	Methods	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10
Industrial	Traditional						
	GRU4Rec	0.0638	0.0542	0.0774	0.0598	0.0999	0.0669
	Caser	0.0618	0.0514	0.0717	0.0555	0.0942	0.0628
	SASRec	0.0790	0.0700	0.0909	0.0748	0.1088	0.0806
	Generative						
	HSTU	0.0927	0.0885	0.1037	0.0918	0.1163	0.0958
	TIGER	0.0852	0.0742	0.1010	0.0807	0.1321	0.0908
	LCRec	0.0915	0.0805	0.1057	0.0862	0.1332	0.0952
	LLM-based						
	BIGRec	0.0931	0.0841	0.1092	0.0907	0.1370	0.0997
	D ³	0.1024	0.0991	0.1213	0.0989	0.1500	0.1082
	S-DPO	0.1032	0.0906	0.1238	0.0991	0.1524	0.1082
	Ours						
	MiniOneRec	0.1143	0.1011	0.1321	0.1084	0.1586	0.1167
Office	Traditional						
	GRU4Rec	0.0629	0.0528	0.0789	0.0595	0.1019	0.0669
	Caser	0.0748	0.0615	0.0865	0.0664	0.1093	0.0737
	SASRec	0.0861	0.0769	0.0949	0.0805	0.1120	0.0858
	Generative						
	HSTU	0.1134	0.1031	0.1252	0.1079	0.1400	0.1126
	TIGER	0.0986	0.0852	0.1163	0.0960	0.1408	0.1002
	LCRec	0.0921	0.0807	0.1048	0.0859	0.1237	0.0920
	LLM-based						
	BIGRec	0.1069	0.0961	0.1204	0.1017	0.1434	0.1091
	D ³	0.1204	0.1055	0.1406	0.1139	0.1634	0.1213
	S-DPO	0.1169	0.1033	0.1356	0.1110	0.1587	0.1255
	Ours						
	MiniOneRec	0.1217	0.1088	0.1420	0.1172	0.1634	0.1242

5.1 EVALUATION SETUP

Experiments are carried out on two real-world slices of the Amazon Review dataset (Hou et al., 2024), namely *Office* and *Industrial*. To measure top- K recommendation accuracy, we follow standard practice and compute Hit Rate (HR@ K) as well as Normalized Discounted Cumulative Gain (NDCG@ K).

5.2 SCALING

We demonstrate the scaling capabilities of MiniOneRec, where the convergence loss decrease consistently as the model size increases. As Figure 1 shows, under the generative recommendation paradigm, there is a clear correlation between model size and loss. As the LLM scale increases, the convergence loss continues to decrease. Furthermore, Figure 3 tracks the evaluation loss on the SFT training set, recorded every 0.5 epoch. It is evident that models with larger parameter counts maintain lower evaluation losses throughout nearly the entire training process and converge more rapidly. Such a superior scaling effect demonstrates the potential of generative recommenders as the next-generation recommendation models.

5.3 PERFORMANCE COMPARISON

Our baselines contain three categories: (1) Traditional recommendation models, including GRU4Rec (Hidasi et al., 2016), Caser (Tang and Wang, 2018), SASRec (Kang and McAuley, 2018); (2)

Generative recommendation models: HSTU (Zhai et al., 2024), TIGER (Rajput et al., 2023), LC-Rec (Zheng et al., 2024); (3) LLM-based recommendation models, including BigRec (Bao et al., 2023), D³ (Bao et al., 2024), S-DPO (Chen et al., 2024b).

We evaluate MINIONEREC on the *Industrial* and *Office* benchmarks and summarize the outcomes in Table 1. Two key insights stand out:

- **Utility of LLM World Knowledge.** Recommenders powered by LLMs, such as BIGRec and D³, clearly surpass traditional systems like GRU4Rec and Caser, showing that the broad knowledge embedded in LLMs translates into better recommendation accuracy.
- **Effectiveness of MiniOneRec.** MiniOneRec goes a step further: by aligning the full generation process with the task objective and using reinforced preference optimization during RL, it consistently outperforms previous generative solutions across most reported metrics. Moreover, by operating in the compact SID space rather than on verbose textual titles, MiniOneRec requires substantially fewer context tokens and yields faster inference, translating into lower latency and smaller memory footprints at serving time.

5.4 TRANSFERABILITY

We evaluate the out-of-distribution (OOD) robustness of MiniOneRec through an experiment we call *SID pattern discovery*. The model is trained exclusively on the *Industrial* domain and then deployed, without any further tuning, to the never-seen *Office* domain. Prior work (Jin et al., 2025; Yue et al., 2025; Yoshihara et al., 2025; Cheng et al., 2025) suggests that SFT may overfit to the source domain and hurt transfer, so we add an RL-only variant, MiniOneRec-w/ RL-OOD, that skips SFT entirely to emphasize generalization. The comparison involves four systems: (1) GRU4Rec, trained and tested inside the *Office* domain; (2) Qwen-Text, which represents user histories as plain sentences and predicts the next item by its title with no additional fine-tuning; (3) Qwen-SID, which encodes the same history with SID tokens and produces the next SID with no additional fine-tuning; (4) MiniOneRec-w/ RL-OOD, trained via GRPO on *Industrial* only and evaluated on *Office*, to evaluate its OOD (out-of-distribution) performance.

Table 2 reports the outcomes. Qwen-Text performs poorly, while Qwen-SID does noticeably better, indicating that a structured SID vocabulary is easier for an LLM to exploit. Although MiniOneRec-w/ RL-OOD falls short of the full MiniOneRec on in-domain scores, its reinforcement-only training grants excellent transfer, achieving competitive accuracy on the unseen catalog. Despite the substantial domain shift and possible semantic drift among SIDs, MiniOneRec successfully uncovers reusable interaction patterns, underscoring the framework’s promise for cross-domain recommendation.

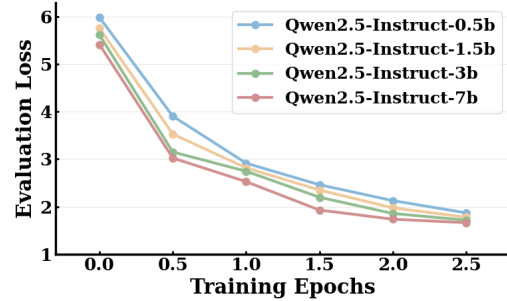


Figure 3: Evaluation loss vs. SFT training epoch

5.5 ABLATION STUDY

To validate the effectiveness of each component in the MiniOneRec framework, we compare it with the following alternative approaches.

5.5.1 ALIGNING STRATEGY

To isolate the impact of each component, we compare the full MiniOneRec with three pared-down variants. (1) MINIONEREC-W/O ALIGN: removes any language-SID alignment and treats recommendation purely as a SID-to-SID task. (2) MINIONEREC-W/ SFTALIGN: keeps the alignment objective during SFT stage only, while RL uses SID data alone. (3) MINIONEREC-W/ RLALIGN: SFT relies solely on SID supervision, and the alignment tasks are introduced later in the RL stage.

Table 2: Performance of MiniOneRec and its variants on completely unseen recommendation domains

Dataset	Method	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10
Office	GRU4Rec	0.0629	0.0528	0.0789	0.0595	0.1019	0.0669
	Qwen-Text	0.0031	0.0021	0.0044	0.0026	0.0057	0.0030
	Qwen-SID	0.0300	0.0214	0.0456	0.0282	0.0733	0.0373
	MiniOneRec-w/ RL-OOD	0.0553	0.0433	0.0691	0.0489	0.0892	0.0553

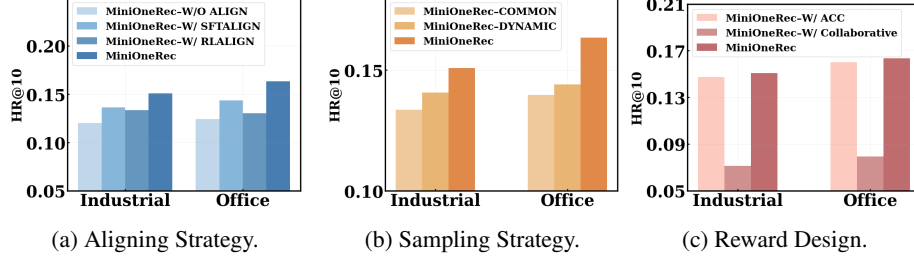


Figure 4: Study on the effectiveness of MiniOneRec’s individual components. Figure 4a examines model performance under different alignment strategies; Figure 4b investigates various sampling strategies; Figure 4c evaluates the impact of alternative reward designs.

Figure 4a summarizes the findings. The complete MiniOneRec, which maintains alignment throughout the whole pipeline, delivers the highest scores on every metric. The MINIONEREC-W/O ALIGN performs worst, indicating that grounding SID generation in world knowledge is essential.

5.5.2 SAMPLING STRATEGY

We study how different roll-out methods affect MiniOneRec by switching only the trajectory generator while keeping everything else fixed: (1) MINIONEREC-COMMON relies on a plain Top- k decoder to produce exactly the required number of paths. (2) MINIONEREC-DYNAMIC follows our two-step sampler: it first draws one and a half times the budget, then retains as many unique items as possible for RL. The full model adopts beam search with width 16.

As plotted in Figure 4b, the complete MiniOneRec delivers the highest accuracy while using roughly two-thirds of the samples needed by the dynamic variant, demonstrating that beam search is the most cost-efficient choice among the tested strategies.

5.5.3 REWARD DESIGN

Three variants are compared: (1) MINIONEREC-W/ ACC that relies solely on a binary correctness signal; (2) MINIONEREC-W/ COLLABORATIVE that replaces the ranking term with logits taken from a frozen SASRec model so as to supply collaborative cues;

As shown in the Figure 4c, the full MiniOneRec achieves the best overall performance. To be noticed, injecting the collaborative reward information into the RL process instead led to significant degradation. We hypothesize that, this stems from reward hacking: as recommendation accuracy declines, the reward continues to increase, revealing a misalignment between this collaborative reward signal and the true objective.

Table 3: Performance of MiniOneRec initialized with pre-trained weights versus random initialization across two domains

Datesets	Methods	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10
Industrial	MiniOneRec-scratch	0.0757	0.0672	0.0891	0.0726	0.1134	0.0804
	MiniOneRec	0.1125	0.0988	0.1259	0.1046	0.1546	0.1139
Office	MiniOneRec-scratch	0.0959	0.0855	0.1057	0.0896	0.1196	0.0941
	MiniOneRec	0.1217	0.1088	0.1420	0.1172	0.1634	0.1242

5.6 PRE-TRAINED LLM IMPACT

To isolate the impact of pre-trained LLMs on generative recommendation, we instantiate two variants of MiniOneRec: one initialized from a general-purpose pre-trained LLM and the other trained from scratch with random weights. As shown in Table 3, Experiments on two benchmark datasets reveal a consistent pattern: the model that starts from pre-trained weights significantly outperforms its randomly initialized counterpart. We conjecture that (i) the general reasoning ability acquired during large-scale language pre-training allows the model to cast the next-SID prediction task as a problem of pattern discovery, as discussed in Section 5.4, and (ii) the factual knowledge already encoded in the LLM offers a head start in understanding the real-world semantics behind each SID, part of which can be transferred to the recommendation domain.

6 CONCLUSION

This report introduces **MiniOneRec**, the first fully open-source generative recommendation framework, which to the best of our knowledge provides an end-to-end workflow spanning SID construction, SFT, and recommendation-oriented RL. By systematically validating scaling laws on public benchmarks, we demonstrate that larger generative recommenders achieve lower training and evaluation losses than their smaller counterparts, confirming the parameter–efficiency advantage of the SID-based paradigm over traditional embedding-centric models. Building on this insight, we introduce post-training techniques: (i) full-process SID alignment, which embeds SID tokens into the model vocabulary and imposes auxiliary alignment tasks across both SFT and RL stages, and (ii) reinforced preference optimization, which combines constrained decoding, beam-based sampling, and hybrid reward design. Extensive experiments on Amazon Review show that MiniOneRec consistently surpasses strong sequential, generative, and LLM-based baselines, while maintaining a lean post-training footprint.

Looking forward, we will keep maintaining and extending the MiniOneRec codebase. A public roadmap will guide future developments, and we warmly welcome community contributions. Planned updates include new datasets, more advanced tokenisation schemes, larger backbone models, and enhanced training pipelines, ensuring that MiniOneRec remains a solid reference platform for research and practice in large-scale generative recommendation.

REFERENCES

- Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *CoRR*, abs/2502.18965, 2025.
- Yuxiang Wang, Xiao Yan, Chi Ma, Mincong Huang, Xiaoguang Li, Lei Yu, Chuan Liu, Ruidong Han, He Jiang, Bin Yin, Shangyu Chen, Fei Jiang, Xiang Li, Wei Lin, Haowei Han, Bo Du, and Jiawei Jiang. Mtgrboost: Boosting large-scale generative recommendation models in meituan. *CoRR*, abs/2505.12663, 2025.
- Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Shen Li, Yanli Zhao, Yuchen Hao, Yantao Yao, Ellie Dingqiao Wen, Jongsoo Park, Maxim Naumov, and Wenlin Chen. Wukong: Towards a scaling law for large-scale recommendation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- Junguang Jiang, Yanwen Huang, Bin Liu, Xiaoyu Kong, Ziru Xu, Han Zhu, Jian Xu, and Bo Zheng. Large language models are universal recommendation learners. *CoRR*, abs/2502.03041, 2025.
- Jie Zhu, Zhifang Fan, Xiaoxie Zhu, Yuchen Jiang, Hangyu Wang, Xintian Han, Haoran Ding, Xinmin Wang, Wenlin Zhao, Zhen Gong, Huizhi Yang, Zheng Chai, Zhe Chen, Yuchao Zheng, Qiwei Chen, Feng Zhang, Xun Zhou, Peng Xu, Xiao Yang, Di Wu, and Zuotao Liu. Rankmixer: Scaling up ranking models in industrial recommenders. *CoRR*, abs/2507.15551, 2025.
- Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. HLLM: enhancing sequential recommendations via hierarchical large language models for item and user modeling. *CoRR*, abs/2409.12740, 2024a.
- Sunhao Dai, Jiakai Tang, Jiahua Wu, Kun Wang, Yuxuan Zhu, Bingjun Chen, Bangyang Hong, Yu Zhao, Cong Fu, Kangle Wu, Yabo Ni, Anxiang Zeng, Wenjie Wang, Xu Chen, Jun Xu, and See-Kiong Ng. Onepiece: Bringing context engineering and reasoning to industrial cascade ranking system. *CoRR*, abs/2509.18091, 2025.
- 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. Language models are few-shot learners. In *NeurIPS*, 2020.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shutong Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, 2024.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur  lien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozi  re, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gr  goire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, J  lmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aur  lien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, 2025.

- Wang-Cheng Kang and Julian J. McAuley. Self-attentive sequential recommendation. In *ICDM*, 2018.
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *ICLR*, 2016.
- Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*, 2018.
- Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 2021.
- Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Trans. Inf. Syst.*, 2020.
- Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. ACM, 2018.
- Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*. ACM, 2016.
- Ruoxi Wang, Rakesh Shivanna, Derek Zhiyuan Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: improved deep & cross network and practical lessons for web-scale learning to rank systems. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 2021.
- Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: A factorization-machine based neural network for CTR prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. ijcai.org, 2017.
- Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, Changqing Qiu, Jiaqi Zhang, Xu Zhang, Zhiheng Yan, Jingming Zhang, Simin Zhang, Mingxing Wen, Zhaojie Liu, Kun Gai, and Guorui Zhou. QARM: quantitative alignment multi-modal recommendation at kuaishou. *CoRR*, abs/2411.11739, 2024.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30, 2022.
- 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. 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*, 2023.
- Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 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*. IEEE, 2024.
- Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. Tokenrec: Learning to tokenize ID for llm-based generative recommendation. *CoRR*, abs/2406.10450, 2024.
- Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, Weifeng Ding, Wuchao Li, Xinchen Luo, Xingmei Wang, Zexuan Cheng, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Di Wang, Dongxue Meng, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang,

- Guowang Zhang, Han Li, Hengrui Hu, Hezheng Lin, Hongtao Cheng, Hongyang Cao, Huanjie Wang, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Liao Yu, Qiang Wang, Qidong Zhou, Shengzhe Wang, Shihui He, Shuang Yang, Shujie Yang, Sui Huang, Tao Wu, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfan Wu, Yunfeng Zhao, and Zhanyu Liu. Onerec technical report. *CoRR*, abs/2506.13695, 2025a.
- Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, Pengfei Zheng, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Ruiming Tang, Shiyao Wang, Shujie Yang, Tao Wu, Wuchao Li, Xinchun Luo, Xingmei Wang, Yi Su, Yunfan Wu, Zexuan Cheng, Zhanyu Liu, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Chenglong Chu, Di Wang, Dongxue Meng, Dunju Zang, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Honghui Bao, Hongyang Cao, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Qiang Wang, Qidong Zhou, Rongzhou Zhang, Shengzhe Wang, Shihui He, Shuang Yang, Siyang Mao, Sui Huang, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yang Zhou, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfeng Zhao, Zhixin Ling, and Ziming Li. Onerec-v2 technical report. *CoRR*, abs/2508.20900, 2025b.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, 2024.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiushi Chen, and Julian J. McAuley. Bridging language and items for retrieval and recommendation. *CoRR*, 2024.
- Yupeng Hou, An Zhang, Leheng Sheng, Zhengyi Yang, Xiang Wang, Tat-Seng Chua, and Julian J. McAuley. Generative recommendation models: Progress and directions. In *WWW (Companion Volume)*, pages 13–16. ACM, 2025.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- Chao Feng, Wuchao Li, Defu Lian, Zheng Liu, and Enhong Chen. Recommender forest for efficient retrieval. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and Zhenhua Dong. EAGER: two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3245–3254. ACM, 2024.
- Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *J. Artif. Intell. Res.*, 1996.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, 2017.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. 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*, 2023.
- Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. In *NeurIPS*, 2024b.
- Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, and Xiang Wang. Llara: Aligning large language models with sequential recommenders. *CoRR*, abs/2312.02445, 2023.

- Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. Language representations can be what recommenders need: Findings and potentials. In *ICLR*, 2025.
- Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference 2024*, 2024.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. Decoding matters: Addressing amplification bias and homogeneity issue in recommendations for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Association for Computational Linguistics, 2024.
- Junfei Tan, Yuxin Chen, An Zhang, Junguang Jiang, Bin Liu, Ziru Xu, Han Zhu, Jian Xu, Bo Zheng, and Xiang Wang. Reinforced preference optimization for recommendation. *arXiv preprint arXiv:2510.12211*, 2025.
- Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Fuli Feng, Xiangnan He, and Qi Tian. A bi-step grounding paradigm for large language models in recommendation systems. *CoRR*, abs/2308.08434, 2023.
- Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. RL is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms, 2025.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *CoRR*, abs/2504.13837, 2025.
- Hiroshi Yoshihara, Taiki Yamaguchi, and Yuichi Inoue. A practical two-stage recipe for mathematical llms: Maximizing accuracy with SFT and efficiency with reinforcement learning. *CoRR*, abs/2507.08267, 2025.
- Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jian-shu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Revisiting reinforcement learning for LLM reasoning from A cross-domain perspective. *CoRR*, abs/2506.14965, 2025.

A ALIGNMENT PROMPTS

A.1 OVERVIEW

In this section we describe how we **align the world knowledge of a LLM with the SID space** so that the model can be directly used for generative recommendation.

We append a three-layer codebook, each layer containing 256 unique SIDs, to the original tokenizer vocabulary. The inserted codes are treated as indivisible tokens, enabling the LLM to read or write SID sequences without any sub-word splitting. Then we introduce a set of auxiliary *Alignment Tasks* to bridge the semantic gap between the textual vocabulary and the newly added SIDs. A specialized *Recommendation Tasks* further teach the model to predict the SID of the next item given an user history.

A.2 RECOMMENDATION TASKS

1. **Generative Retrieval.** The main task ofor generative recommendation. The LLM receives a chronologically ordered SID sequence that represents the user’s recent interactions, together with an explicit instruction such as “Recommend the next item.”, and is asked to predict the SID of the next item the user might engage with. A sample prompt is shown in the Figure 5.

Input:
The user has interacted with items <a_13><b_197><c_1>, <a_52><b_17><c_113>, <a_13><b_201><c_34> in chronological order. Can you predict the next possible item that the user may expect?
Response:
<a_13><b_72><c_149>

Figure 5: Generative Retrieval Prompt.

2. **Asymmetric Item Prediction.** (a) Given a textual user history, predict the SID of the next item; (b) given a SID-only history, generate the textual title of the next item. Sample prompts are shown in the Figure 6 and Figure 7.

Input:
The user has interacted with items 'Kreg SML-C150-100 Pocket Screws 1-1/2-Inch, 8 Coarse, Washer-Head, 100-Count', '3M Flap Disc 566A, T29, 4-1/2" Diameter, 40 Grit, 5/8"-11 Thread (Pack of 1)' in chronological order. Can you predict the next possible item that the user may expect?
Response:
<a_104><b_60><c_152>

Figure 6: Asymmetric Item Prediction Prompt1.

Input:
The user has interacted with items <a_71><b_44><c_249>, <a_71><b_114><c_136>, <a_67><b_244><c_35> in chronological order. Can you predict the next possible item’s title that the user may expect?
Response:
How the Grinch Stole Christmas!: Mini Edition (Dr Seuss Miniature Edition)

Figure 7: Asymmetric Item Prediction Prompt2.

A.3 ALIGNMENT TASKS

1. **SID–Text Semantic Alignment.** (a) Predict an item’s textual title from its SID; (b) predict the SID from the item’s textual title. Sample prompts are shown in the Figure 11 and Figure 9.

Input:
What is the title of <a_24><b_141><c_73> ?
Response:
Oral-B Deep Sweep Toothbrush

Figure 8: SID–Text Semantic Alignment Prompt1.

Input:
Which item has the title *Nashua Stretch & Seal Self-Fusing Silicone Tape* ?
Response:
<a_202><b_202><c_29>

Figure 9: SID–Text Semantic Alignment Prompt2.

2. **Item Description Reconstruction.** To ground SIDs in richer semantics, we ask the model to generate the item description from a single SID and, conversely, infer the SID from the description. We perform this task only during the SFT stage because the description space is large and diverse. Sample prompts is shown in the Figure 10.

Input:
An item can be described as follows: *Relay/CAPACITOR hard start kit 500% incr starting torque, increased compressor starting torque. Designed for use on permanent split CAPACITOR single phase a/c and heat pump systems. Multiple use includes: room air conditioner units, residential or commercial psc AC units.. the country of origin is China.* Which item is it describing?
Response:
<a_113><b_16><c_2>

Figure 10: Item Description Reconstruction Prompt.

3. **User Preference Summarization.** Given a sequence of SIDs, the model produces a short natural–language profile that summarizes the user’s interests. Because the raw dataset lacks explicit preference labels, we employ DEEPSEEK (DeepSeek-AI et al., 2024) to extract summaries from the item’s meta data and users’ textual reviews and use them as pseudo labels. This task, too, is restricted to the SFT stage due to the open-ended output space.

Input:
The user has interacted with items <a_39><b_41><c_1>, <a_39><b_28><c_16>, <a_67><b_142><c_35> in chronological order. Can you summarize the user’s preference?
Response:
Based on the items purchased, this user has a clear preference for durable, high-performance, and practical supplies for hands-on tasks..... electrical work, and general repairs, necessitating dependable personal protective equipment and precision tools.

Figure 11: User Preference Summarization Prompt.

B DATASETS

We evaluate our model on two Amazon Review subsets: Amazon Review subsets (*Industrial_and_Scientific* and *Office_Products*). To keep computational costs affordable, we follow the trimming strategy used in (Bao et al., 2024). The main steps are: (1) remove users and items with fewer than five interactions; (2) for *Toys_and_Games*, keep events from October 2016 to November 2018; (3) for *Industrial_and_Scientific*, which is smaller, keep all events between October 1996 and November 2018; (4) truncate every user history to at most ten items; (5) finally, split each dataset chronologically into training, validation and test sets with an 8:1:1 ratio. Key statistics of the resulting training splits are reported in Table 4.

Table 4: Statistics of datasets.

Datasets	Industrial	Office
Items	3,685	3,459
Train	3,6259	3,8924
Valid	4,532	4,866
Test	4,533	4,866