Rethinking LLM-Based Recommendations: A Personalized Query-Driven Parallel Integration

Donghee Han **Hwanjun Song**[†] **Mun Yong Yi**[†]
KAIST, Republic of Korea
{handonghee, songhwanjun, munyi}@kaist.ac.kr
[†]Corresponding authors

Abstract

Recent studies have explored integrating large language models (LLMs) into recommendation systems but face several challenges, including training-induced bias and bottlenecks from serialized architecture. To effectively address these issues, we propose a **Query-to-Recommendation**, a **parallel recommendation framework** that decouples LLMs from candidate pre-selection and instead enables direct retrieval over the entire item pool. Our framework **connects LLMs and recommendation models** in a parallel manner, allowing each component to independently utilize its strengths without interfering with the other. In this framework, LLMs are utilized to generate feature-enriched item descriptions and personalized user queries, allowing for capturing diverse preferences and enabling rich semantic matching in a zero-shot manner. To effectively combine the complementary strengths of LLM and collaborative signals, we introduce an adaptive reranking strategy. Extensive experiments demonstrate an improvement in performance up to 57%, while also improving the novelty and diversity of recommendations.

1 Introduction

Recommendation systems play a crucial role in delivering personalized content and service across various domains. As the demand for more accurate and diverse recommendations continues to increase, the integration of large language models (LLM) has emerged as a promising advancement (Wu et al., 2024; Zhao et al., 2024). LLMs possess extensive knowledge and exhibit remarkable abilities in understanding and generating text (Yang et al., 2024), enabling new opportunities to improve recommendation systems beyond the traditional collaborative filtering (CF) and content-based methods. Recently, numerous studies have been proposed to leverage LLM in recommendation systems (Ramos et al., 2024; Zhang et al., 2025a; Lu et al., 2024).

To harness the capabilities of LLMs for recommendation, two major research paradigms have emerged. The first line of works utilize LLMs as generative predictors, typically fine-tuning them on next-item prediction tasks. These methods generate textual representations (e.g., item titles) based on user histories and use them as queries for retrieving candidate items (Bao et al., 2025; Li et al., 2023b). While this approach leverages LLMs' generation strength, it requires fine-tuning, which introduces bias to train dataset and reduces diversity of recommendation. Additionally, relying solely on item titles can limit the expressive capacity of LLMs, limiting their ability to align nuanced user preferences with rich item characteristics.

The second line of research focuses on reranking candidate items directly within the LLM prompt. In this setting, a separate candidate retrieval model selects a subset of items, which are then presented to the LLM for scoring or reranking (Hou et al., 2024; Kim et al., 2024; Bao et al., 2023). These approaches typically leverage CF-based recommendation models such as SASRec (Kang and McAuley, 2018) for candidate selection (Yang et al., 2023), and various techniques have been proposed to align LLMs with CF signals (Kim et al., 2024; Dong et al., 2025).

However, these approaches inherently depend on the performance of the candidate selector, and they limit the ability of LLMs to fully leverage their distinctive capabilities across the entire item pool. The information handled by the CF-based model and the LLM is fundamentally different, and the current serialized architecture can limit performance due to the misalignment between these two components. This misalignment also becomes a bottleneck that hinders the utilization of the diverse knowledge embedded in LLMs for recommendation.

In this study, we focus on addressing two key challenges in LLM-based recommendation: the bias introduced by fine-tuning LLMs on recommen-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8395–8419
November 4-9, 2025

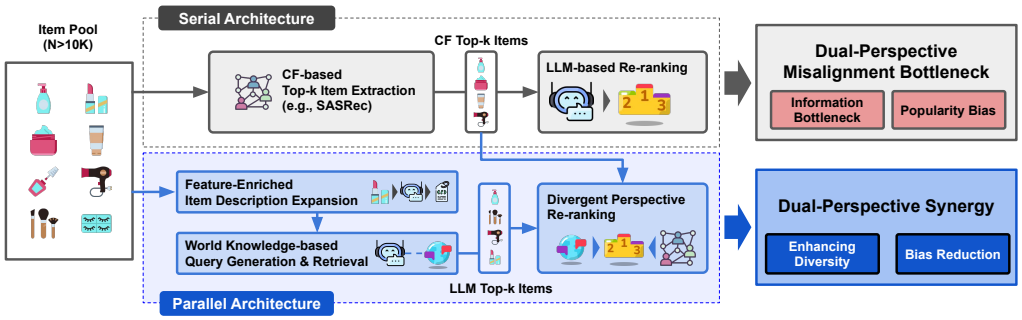


Figure 1: Comparison between the traditional serialized pipeline and the proposed parallel approach.

dation datasets, and the bottleneck caused by the serialized architecture that utilizes CF-based models as candidate selectors. Motivated by the fact that CF-based models and LLMs handle fundamentally different types of information, we propose a parallel framework that integrates both components without entangling their processes. Our approach enables the generation of enriched, feature-aware queries that allow LLMs to effectively leverage their broad and diverse knowledge for recommendation.

Fig.1 illustrates the difference between the traditional serialized approach and our proposed parallel approach. Our proposed parallel architecture maximizes the utilization of the LLM's world knowledge and creates synergy by effectively merging the two complementary perspectives. We argue that LLMs can offer a complementary perspective to CF-based recommendation models. To fully leverage their potential, LLMs need direct access to the global item pool; however, existing approaches that fine-tune LLMs to generate item titles hinder the utilization of the diverse knowledge of LLMs and introduce biases from the training dataset..

In our framework, LLMs and CF-based models independently retrieve top-k items from the entire candidate pool. By decoupling LLMs from preselected candidates, our method maximizes LLMs' ability and creates synergy between complementary perspectives. To utilize LLMs' rich world knowledge, we introduce item description expansion and personalized query generation where the LLM generates feature-enriched item description and personalized user queries grounded in user histories and item characteristics. Furthermore, we design an adaptive reranking strategy that aggregates the outputs of LLMs and CF-based models while maintaining their independency. Unlike existing approaches that jointly train multiple models, our method dynamically balances contributions.

Our method not only improves the performance of the recommendation, but also promotes greater diversity and adaptability, effectively addressing the limitations of existing LLM-based recommendation systems. The training-free nature of our framework allows for seamless incorporation of new items based on their key features without additional model updates, and it can also handle large, dynamically changing item pools. This advantage is particularly beneficial in real-world e-commerce scenarios, where large volumes of new items are continuously introduced and updated. Consequently, our design significantly reduces both infrastructure costs and maintenance overhead, making the system highly scalable for rapidly evolving recommendation environments.

Our contributions in this work are as follows.

- We **introduce a parallel architecture** for recommendation that combines LLM-based retrieval and CF-based models, maximizing the complementary strengths of each component.
- Our method **outperforms state-of-the-art methods** and achieved up to 57% improvement in performance over traditional CF-based models, without requiring additional LLM fine-tuning.
- Through our analysis, we demonstrate that our approach effectively **addresses bias and diversity issues** in recommendations.

2 Related Work

2.1 Traditional Recommendation Systems

Sequential recommendation has emerged as a prominent paradigm, leveraging users' recent interactions to predict the next likely item of interest. Deep learning-based models, such as SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019), have demonstrated strong performance in capturing complex sequential user behaviors, effectively modeling temporal

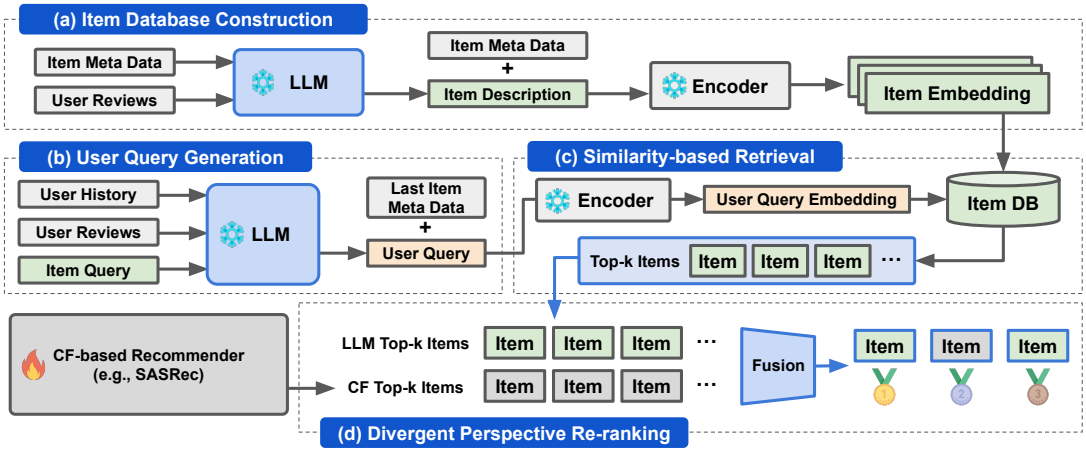


Figure 2: Overview of QUEREC.

dependencies to improve the precision of the recommendation. Recommendation systems that rely solely on item IDs fail to capture the key features of the items. To address this limitation, various sequential recommendation models have been proposed that incorporate additional information (Xie et al., 2022; Yuan et al., 2023).

2.2 LLM-based Recommendation Systems

With the rapid advancement of LLMs, there has been a growing interest in incorporating them into recommendation systems (Wu et al., 2024). Early LLM-based recommendation approaches fine-tuned pre-trained models such as T5 (Raffel et al., 2020) to generate item recommendations by providing item or user IDs and output item IDs in a text format (Geng et al., 2022; Li et al., 2023c; Wang et al., 2024b,a). However, these approaches had limitations in fully leveraging the rich knowledge embedded in LLMs.

Another direction in LLM-based recommendation research explores natural language prompts instead of item and user IDs, allowing LLMs to leverage their extensive knowledge and zero-shot capabilities. Early studies fine-tuned LLMs (Bao et al., 2023), while later approaches fixed LLM parameters and update only additional components to incorporate CF-signals (Kim et al., 2024; Dong et al., 2025). Recent work introduced zero-shot LLM-based reranking Hou et al. (2024), but these methods remain limited by the reliance on upstream candidate selectors and positional bias.

Recently, other studies proposed fine-tuning LLMs to generate item titles and using the generated titles as search queries for recommendation(Li et al., 2023b; Bao et al., 2025). However, these approaches have limitations in leveraging the diverse knowledge inherent in LLMs. To address these challenges, we propose a method that better aligns with real-world recommendation settings, ensuring ease of deployment and scalability in large-scale item pools.

3 Method: QUEREC Framework

In this section, we introduce QUEREC(Query-to-Recommendation), a novel framework designed to leverage the LLMs through large item pool and to operate in parallel with CF-based models, enabling a dual-perspective that integrates both collaborative signals and language-based knowledge. Fig.2 illustrates the overall process of our method. The proposed method leverages LLMs to generate feature-enriched item descriptions and personalized user queries, retrieving top-k items based on similarity scores between user queries and item descriptions. These LLM-based results are then combined with the top-k items from a CF-based model, effectively fusing insights from both CF-based models and LLMs.

3.1 Item Database Construction

To effectively leverage the world knowledge of LLMs to highlight key features of items and align them with user preferences, we employ LLMs to generate feature-enriched item descriptions that reflect the distinctive features of each item. Fig.2(a) illustrates the process of leveraging an LLM to expand item descriptions, which are then encoded into embeddings to construct an item vector database.

This process involves constructing prompts using item metadata (e.g. title, category, description) and user reviews to provide contextual information to the LLM. We design the prompt to expand item

descriptions into a set of diverse queries, aiming to better reflect users' varied preferences. Given these prompts, the LLM generates ten distinct queries, each capturing different aspects of the item by leveraging its extensive knowledge. Finally, the generated queries are combined with the original item metadata to create a comprehensive item representation, which is then converted into an item embedding using a retrieval encoder and cached for search.

$$q_i = \text{LLM}(p_i, m_i, r_i), \quad d_i = m_i \oplus q_i \tag{1}$$

Equation (1) describes the process of extracting item queries and constructing the complete item representation corpus. Given an item i, we construct a query q_i using the LLM with inputs: item query generation prompt p_i, item metadata m_i, and user reviews r_i. The final textual representation d_i is formed by concatenating the metadata of the item with the generated query. Here, $\oplus$ represents the concatenation operator, integrating both structured metadata and insights derived from LLM to improve the retrieval effectiveness of items.

$$\mathcal{V} = \{\text{Enc}(d_i) \mid d_i \in \mathcal{D}\} \tag{2}$$

Continuing from the previous step, we encode all enriched item descriptions $d_i \in \mathcal{D}$ using a retrieval encoder and cache the resulting embeddings $\mathcal{V}$ for efficient search during recommendation, as shown in Equation (2).

3.2 User Query Generation

To represent the user, we incorporate information from their historical interactions and preferences. This process begins with the construction of prompts that combine user history (including item titles and metadata), user reviews, and previously generated item descriptions to provide rich contextual information. Based on these prompts, the LLM generates ten personalized queries that reflect the user's context and preferences. To emphasize the user's most recent preferences, we append the enriched description of the most recently purchased item to the prompt.

$$q_u = \text{LLM}(p_u, h_u, r_u, d_l), \quad d_u = m_l \oplus q_u \tag{3}$$

Equation (3) describes the process of constructing the textual user representation. Given a user u, we construct a query q_u using the LLM with inputs: user query generation prompt p_u, user history h_u (which includes item titles and metadata), user reviews r_u, and the most recent item description d_l. The final user representation d_u is obtained by concatenating the user history with the generated query and incorporating the last interacted item metadata m_l to better reflect temporal user preferences.

An important aspect of this step is that, instead of using each generated query independently, we aggregate them into a single unified user representation. Similar to the item description expansion, we prompt the LLM to generate multiple personalized queries in order to capture diverse user perspectives and preferences. These queries are not treated separately but are concatenated into a single user representation, which is then used to construct a comprehensive user embedding. The prompt format and generated query examples are described in the Appendix G.

3.3 Similarity-based Retrieval

With the user textual representations and item database obtained from the previous steps, we perform similarity-based retrieval for personalized recommendation :

$$\begin{aligned} v_u &= \text{Enc}(d_u), \quad v_i \in \mathcal{V} \\ s_{u,i} &= \cos(v_u, v_i) \end{aligned} \tag{4}$$

$$\hat{\mathcal{I}}_u = \arg\max_{i \in \mathcal{I}} s_{u,i}, \quad \text{where } |\hat{\mathcal{I}}_u| = k \tag{5}$$

Equation (4) defines the embedding extraction via the pre-trained text encoder Enc and similarity computation between the user embedding v_u and item embeddings in item database $v_i \in \mathcal{V}$. We compute cosine similarity $s_{u,i}$ between the embeddings to measure their relevance. Based on these scores, we select the top-k items $\hat{\mathcal{I}}_u$ for the user query, as shown in Equation (5), which are the most relevant items to the user's preferences.

Our semantic similarity-based retrieval utilizing personalized queries offers richer information compared to traditional grounding methods that rely solely on item titles. It is also more adaptable to updates in the item pool, which occur frequently in real-world recommendation scenarios.

3.4 Divergent Perspective Reranking

To effectively integrate LLM-based semantic insights and CF-based collaborative signals, we propose a divergent perspective reranking method. Initially, similarity scores from each model (LLM and CF) are normalized, ensuring fair influence from each method:

$$\tilde{s}^X_{u,i} = \frac{s^X_{u,i} - \min(s^X_u)}{\max(s^X_u) - \min(s^X_u)}, \quad X \in \{LLM, CF\}. \tag{6}$$

In Equation (6), $s^X_{u,i}$ and $\tilde{s}^X_{u,i}$ denote the original score and final normalized score, between user u and item i computed by model X (LLM-based and CF-based). For each user, we apply min-max normalization over the score set $s^X_u = \{s^X_{u,i} \mid i \in \mathcal{I}_u\}$, where $\mathcal{I}_u$ represents the entire item pool. This normalization maps all user-item relevance scores to the range $[0, 1]$, ensuring that the scores from different models are comparable on the same scale. The process preserves the internal ranking structure of each model while aligning their score magnitudes for reranking.

Subsequently, we calculate the final reranking score using a convex combination (CC) of the normalized scores:

$$s^*_{u,i} = \lambda \tilde{s}^{LLM}_{u,i} + (1 - \lambda)\tilde{s}^{CF}_{u,i}. \tag{7}$$

The initial λ is determined by the validation Hit@10 performance of each model:

$$\lambda_{\text{init}} = \frac{H^{LLM}_{10}}{H^{LLM}_{10} + H^{CF}_{10}}. \tag{8}$$

This method of computing the initial λ is intended to assign a larger weight to the model with higher performance. However, it may be less effective in cases where the two models perform well on entirely different user groups.

Further analysis revealed that the intersection ratio between hit items from QUEREC and CF models varies significantly across datasets and models. In the case of a low intersection ratio with imbalanced weights (e.g., 0.3), interference between the two recommended item lists causes performance to decrease. To address this, we adjust λ by incorporating the intersection ratio:

$$\lambda = \omega \cdot \lambda_{\text{init}} + (1 - \omega) \cdot 0.5. \tag{9}$$

Here, ω is computed as:

$$\omega = \frac{|\mathcal{H}^{\text{LLM}}_{10} \cap \mathcal{H}^{\text{CF}}_{10}|}{|\mathcal{H}^{\text{LLM}}_{10} \cup \mathcal{H}^{\text{CF}}_{10}|}, \tag{10}$$

where $\mathcal{H}^{\text{LLM}}_{10}$ and $\mathcal{H}^{\text{CF}}_{10}$ denote the sets of user sequences whose top-10 lists generated by the LLM-based and CF-based models, respectively, contain the target item.

When the intersection ratio is minimized ($\omega \approx 0$), λ converges to 0.5, reflecting the orthogonality between the two models, and ensuring a balanced influence from both models. This adjustment improves the quality of the recommendation by effectively merging diverse perspectives from the LLM and CF-based method. The analysis related to this approach is explained in Appendix D.

4 Experiment

This section presents the experimental results and analysis that demonstrate the effectiveness of our proposed method.

4.1 Experimental Setup

To evaluate the effectiveness of QUEREC, we integrate existing CF-based recommendation method into our method and compared its performance against existing LLM-based reranking methods. Our goal was to examine whether the proposed parallel architecture would be more effective than the traditional serial reranking approach. For scalability, we employed LLaMA3.2 3B (Dubey et al., 2024) for query generation and the implementation of LLM-based baselines, while for similarity-based retrieval, we leveraged a widely used pre-trained encoder (Li and Li, 2023) without any additional training.

❶ In the first experiment, we evaluated the performance of QUEREC in comparison to existing LLM-based reranking methods when adapting sequential recommendation models. For the existing LLM-based reranking methods, recommendation models are used as candidates selectors similar to the previous study (Yang et al., 2023). ❷ In the second experiment, we compared QUEREC against existing CF-based approaches, T5-based approaches, and retrieval-based approaches that involve fine-tuning LLMs for query generation.

To better reflect real-world recommendation scenarios with a large candidate pool, we evaluated performance over the entire item pool. We adopted two widely used ranking metrics: *Hit Rate (HR)* and *Normalized Discounted Cumulative Gain (NDCG)*, with $k \in \{5, 10\}$. More details on the experimental settings and implimentation are provided in the appendix and online repository[1].

[1] https://github.com/venzino-han/QueRec2025

Model	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
SASRec	0.0328	0.0179	0.0499	0.0234	0.0561	0.0323	0.0872	0.0423	0.0628	0.0354	0.0915	0.0447
+ LLMRank ICL	0.0201	0.0099	0.0341	0.0144	0.0316	0.0156	0.0555	0.0233	0.0410	0.0204	0.0636	0.0277
+ LLMRank Seq	0.0187	0.0088	0.0328	0.0134	0.0279	0.0134	0.0506	0.0208	0.0388	0.0190	0.0624	0.0267
+ LLMRank Recent	0.0198	0.0099	0.0335	0.0143	0.0304	0.0154	0.0525	0.0225	0.0415	0.0213	0.0639	0.0286
+ TALLRec (w/o FT)	0.0092	0.0046	0.0172	0.0072	0.0077	0.0038	0.0145	0.0060	0.0236	0.0122	0.0409	0.0178
+ TALLRec (w FT)	0.0036	0.0028	0.0054	0.0033	0.0048	0.0029	0.0085	0.0041	0.0097	0.0069	0.0125	0.0078
+ A-LLMRec	0.0200	0.0097	0.0435	0.0174	0.0063	0.0029	0.0311	0.0109	0.0493	0.0243	0.0911	0.0380
QUEREC (+ SASRec)	**0.0367**	**0.0242**	**0.0571**	**0.0307**	**0.0675**	**0.0462**	**0.1022**	**0.0574**	**0.0856**	**0.0594**	**0.1189**	**0.0701**
Improvement	11.89%	35.20%	14.43%	31.20%	20.32%	43.03%	17.20%	35.70%	36.31%	67.80%	29.95%	56.82%
DIF-SR	0.0360	0.0197	0.0542	0.0256	0.0576	0.0336	0.0901	0.0442	0.0700	0.0404	0.0995	0.0499
+ LLMRank ICL	0.0210	0.0104	0.0365	0.0154	0.0323	0.0159	0.0545	0.0231	0.0393	0.0197	0.0643	0.0278
+ LLMRank Seq	0.0185	0.0088	0.0341	0.0138	0.0289	0.0139	0.0509	0.0210	0.0384	0.0189	0.0626	0.0268
+ LLMRank Recent	0.0204	0.0102	0.0356	0.0151	0.0318	0.0161	0.0532	0.0230	0.0433	0.0223	0.0656	0.0296
+ TALLRec (w/o FT)	0.0168	0.0095	0.0325	0.0145	0.0239	0.0166	0.0422	0.0224	0.0307	0.0173	0.0577	0.0260
+ TALLRec (w FT)	0.0182	0.0150	0.0188	0.0152	0.0182	0.0150	0.0188	0.0152	0.0191	0.0115	0.0316	0.0156
+ A-LLMRec	0.0240	0.0122	0.0449	0.0191	0.0093	0.0046	0.0328	0.0121	0.0362	0.0151	0.0896	0.0328
QUEREC (+ DIF-SR)	**0.0371**	**0.0247**	**0.0590**	**0.0318**	**0.0699**	**0.0472**	**0.1032**	**0.0579**	**0.0885**	**0.0615**	**0.1228**	**0.0726**
Improvement	3.06%	25.38%	8.86%	24.22%	21.35%	40.48%	14.54%	31.00%	26.43%	52.23%	23.42%	45.49%
ELMRec	0.0492	0.0414	0.0569	0.0437	0.0610	0.0503	0.0729	0.0540	0.0706	0.0616	0.0749	0.0623
+ LLMRank ICL	0.0229	0.0128	0.0312	0.0155	0.0290	0.0157	0.0409	0.0196	0.0296	0.0169	0.0367	0.0192
+ LLMRank Seq	0.0154	0.0084	0.0279	0.0124	0.0236	0.0128	0.0355	0.0166	0.0279	0.0162	0.0366	0.0190
+ LLMRank Recent	0.0167	0.0091	0.0272	0.0125	0.0236	0.0129	0.0361	0.0169	0.0275	0.0160	0.0346	0.0183
+ TALLRec (w/o FT)	0.0214	0.0122	0.0277	0.0142	0.0221	0.0128	0.0309	0.0156	0.0290	0.0173	0.0352	0.0194
+ TALLRec (w FT)	0.0151	0.0086	0.0207	0.0104	0.0084	0.0047	0.0112	0.0056	0.0259	0.0155	0.0315	0.0173
+ A-LLMRec	0.0007	0.0003	0.0026	0.0009	0.0000	0.0000	0.0004	0.0001	0.0005	0.0002	0.0012	0.0004
QUEREC (+ ELMRec)	**0.0589**	**0.0455**	**0.0746**	**0.0506**	**0.0816**	**0.0625**	**0.1060**	**0.0704**	**0.1092**	**0.0777**	**0.1350**	**0.0860**
Improvement	19.72%	9.90%	31.11%	15.79%	33.77%	24.25%	45.40%	30.37%	54.67%	26.14%	80.24%	38.04%

Table 1: Evaluation results in the LLM-recommender cooperation scenario. The best and second-best results for each metric are highlighted in bold and underlined, respectively. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

4.2 Main Results

In this subsection, we present key results comparing our method with existing baselines.

4.2.1 LLM-recommender Cooperation Scenario

The first experiment evaluated the effectiveness of our proposed parallel architecture. The results of this experiment are presented in Table 1. We observed that existing LLM-based reranking methods lead to a decrease in performance. This decline seems to be due to the fact that the selected candidate items from the CF-based model are very similar to each other, which may confuse the LLM during reranking. Although A-LLMRec, which uses CF-based embeddings, outperformed other LLM-based reranking methods, it still underperformed compared to our method. These results suggest that learning CF-signals through natural language-based training remains a challenging task, even when the LLM is trained on candidate items from CF-based models.

In contrast, QUEREC achieved an average performance improvement of 31% and a maximum improvement of 57%. In particular, QUEREC does not require additional training when integrated into an existing recommendation system, highlighting its epandability and ease of deployment. These results demonstrate that QUEREC can effectively leverage the knowledge inherent in LLM, thereby providing differentiated recommendations from CF-based models.

4.2.2 Compare to Traditional Recommenders

In the second experiment, we compared the performance of QUEREC against traditional sequential recommendation models and LLM-based retrieval methods. The results presented in Table 2 demonstrate that QUEREC outperforms existing recommendation models. Furthermore, QUEREC significantly outperformed existing LLM fine-tuning methods such as GPT4Rec and BIGRec, which generate item-titles and use them as queries. These results suggest that approaches based on fine-tuning may introduce biases induced by datasets and potentially compromise the inherent generalization capabilities of LLMs.

4.3 In-depth Analysis

In this subsection, we analyze the distinguishing advantages of the proposed method.

4.3.1 Ablation Study

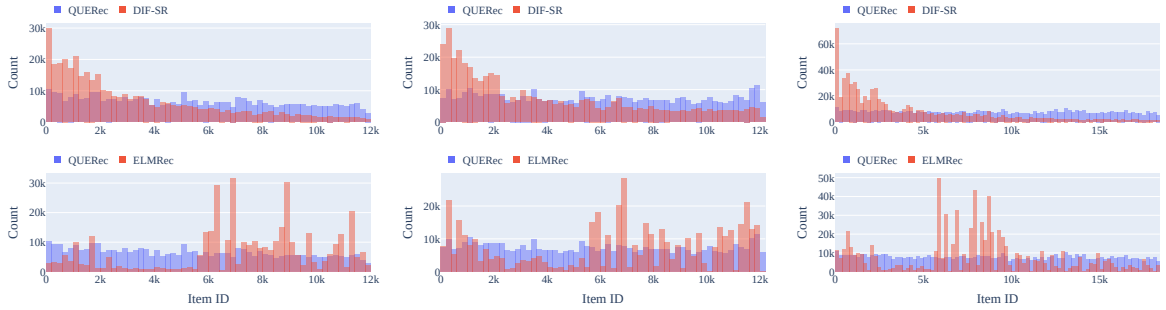
We conducted an ablation study to assess the contribution of each component in our approach. As shown in Table 3, we evaluated performance changes when each of the key components was

Method	Model	Sports				Beauty				Toys			
		H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
CF-based	Caser	0.0116	0.0072	0.0194	0.0097	0.0205	0.0131	0.0347	0.0176	0.0166	0.0107	0.0270	0.0141
	GRU4Rec	0.0129	0.0086	0.0204	0.0110	0.0164	0.0099	0.0283	0.0137	0.0097	0.0059	0.0176	0.0084
	HGN	0.0189	0.0120	0.0313	0.0159	0.0325	0.0206	0.0512	0.0266	0.0321	0.0221	0.0497	0.0277
	SASRec	0.0328	0.0179	0.0499	0.0234	0.0561	0.0323	0.0872	0.0423	0.0628	0.0354	0.0915	0.0447
	BERT4Rec	0.0115	0.0075	0.0191	0.0099	0.0203	0.0124	0.0347	0.0170	0.0116	0.0071	0.0203	0.0099
	FDSA	0.0182	0.0122	0.0288	0.0156	0.0267	0.0163	0.0407	0.0208	0.0228	0.0140	0.0381	0.0189
	S^3-Rec	0.0251	0.0161	0.0385	0.0204	0.0387	0.0244	0.0647	0.0327	0.0443	0.0294	0.0700	0.0376
	DIF-SR	0.0360	0.0197	0.0542	0.0256	0.0576	0.0336	0.0901	0.0442	0.0700	0.0404	0.0995	0.0499
T5-based	P5	0.0272	0.0169	0.0361	0.0198	0.0503	0.0370	0.0659	0.0421	0.0648	0.0567	0.0709	0.0587
	TIGER	0.0264	0.0181	0.0400	0.0225	0.0454	0.0321	0.0648	0.0384	0.0521	0.0371	0.0712	0.0432
	POD	0.0496	0.0396	0.0576	0.0419	0.0537	0.0395	0.0688	0.0443	0.0691	0.0599	0.0742	0.0610
	RDRec	0.0505	0.0408	0.0596	0.0433	0.0601	0.0461	0.0743	0.0504	0.0723	0.0593	0.0802	0.0605
	ELMRec	0.0492	0.0414	0.0569	0.0437	0.0610	0.0503	0.0729	0.0540	0.0706	0.0616	0.0749	0.0623
Retrieval-based	GPT4Rec	0.0002	0.0001	0.0004	0.0002	0.0004	0.0002	0.0006	0.0003	0.0005	0.0003	0.0011	0.0005
	BIGRec	0.0071	0.0046	0.0115	0.0061	0.0237	0.0169	0.0355	0.0207	0.0324	0.0229	0.0471	0.0276
Ours	**QUEREC**	**0.0589**	**0.0455**	**0.0746**	**0.0506**	**0.0816**	**0.0625**	**0.1060**	**0.0704**	**0.1092**	**0.0777**	**0.1350**	**0.0860**

Table 2: Performance comparison of existing recommendation methods. The best results for each metric are highlighted in bold. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

Method	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
QUEREC	**0.0589**	**0.0455**	**0.0746**	**0.0506**	**0.0816**	**0.0625**	**0.1060**	**0.0704**	**0.1092**	**0.0777**	**0.1350**	**0.0860**
w/o CF-based model	0.0280	0.0188	0.0403	0.0228	0.0501	0.0354	0.0675	0.0410	0.0680	0.0472	0.0937	0.0555
w/o recent item info	0.0274	0.0185	0.0407	0.0228	0.0459	0.0319	0.0622	0.0372	0.0625	0.0434	0.0908	0.0525
w/o item description	0.0248	0.0165	0.0369	0.0204	0.0454	0.0317	0.0625	0.0373	0.0625	0.0423	0.0888	0.0508
w/o user query	0.0095	0.0061	0.0166	0.0084	0.0211	0.0131	0.0325	0.0168	0.0254	0.0162	0.0412	0.0213

Table 3: Evaluation results of ablation study. The best results for each metric are highlighted in bold.



(a) Toys. (b) Beauty. (c) Sports.

Figure 3: Diversity of recommendations across different datasets and models. The x-axis represents the IDs of the recommended items, while the y-axis indicates the frequency of recommendation. The top-20 recommended items for each user were extracted and compared.

removed: (1) CF-based model (2) recent item information, (3) generated item description, and (4) user queries. The absence of user queries resulted in the most significant performance degradation, while excluding the other components also led to notable declines. Another notable finding is that our method outperforms fine-tuned LLM-based approaches (GPT4Rec and BIGRec), even without being combined with a CF-based model. This result is encouraging, as our method does not require any LLM fine-tuning steps.

4.3.2 Diversity of Recommendations

Our proposed method does not rely on additional training, allowing unbiased and more diverse recommendations compared to traditional methods that often reinforce popularity bias through training. As shown in Fig.3, existing methods exhibit skewed item distributions, suggesting that the training process introduces biases into these models. In contrast, our proposed method, produces a more balanced distribution of recommendation, indicating improved diversity.

This characteristic can mitigates filter bubbles and promotes exploratory user experiences. These findings highlight that our method not only incorporates a distinct perspective compared to traditional recommendation models, but also has the potential to generate synergies when integrated with existing approaches. Further analysis is provided in AppendixE.1.

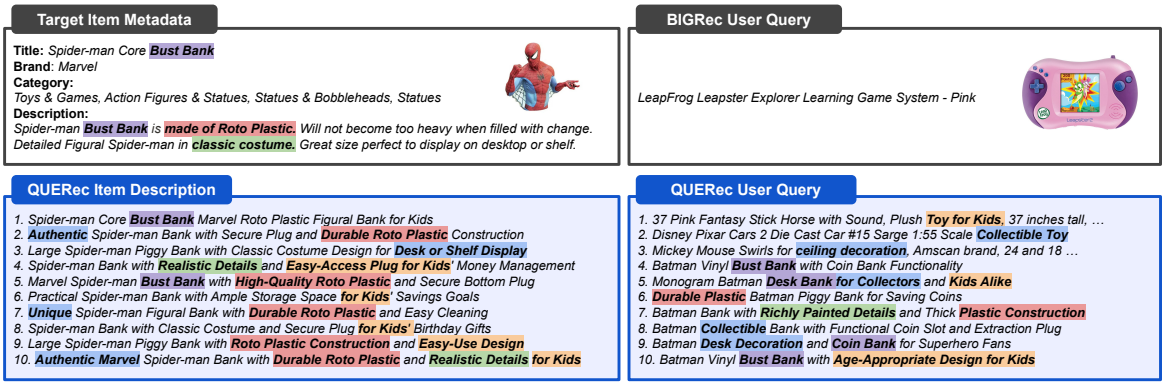


Figure 4: Query quality comparison in the Toys Dataset. Related attributes are highlighted in the same color.

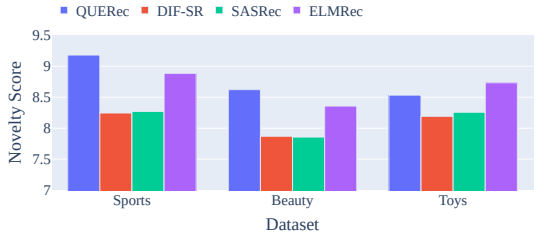


Figure 5: Novelty evaluation in three datasets.

4.3.3 Novelty Evaluation

To examine whether the observed improvement in the diversity of recommendations translates into the recommendation of more novel items, we analyze based on the average novelty scores of items in the Hit@10 results. Novelty is computed as the negative logarithm of the relative item popularity (Zhou et al., 2010), defined as:

$$\text{Novelty}(i) = -\log\left(\frac{\text{freq}(i)}{\sum_{j\in\mathcal{I}}\text{freq}(j)}\right), \quad (11)$$

where $\text{freq}(i)$ denotes the frequency of item i in the training set, and $\mathcal{I}$ is the set of all items.

Fig. 5 presents the average novelty scores of the top-10 recommended items for each method. In most cases, our proposed method outperformed existing approaches, indicating its effectiveness in promoting the recommendation of less popular and more novel items. These results suggest that our proposed method is capable of capturing fundamentally different signals compared to traditional recommendation systems, indicating its potential to deliver personalized recommendations tailored to users with diverse preferences.

4.3.4 Query Quality Analysis

To examine whether the proposed method effectively leverages the broad knowledge encoded in LLMs, we qualitatively compare the generated queries. Existing approaches, which are fine-tuned to produce item-title-style queries, often exhibit dataset-induced biases and struggle to provide clear reasoning for recommendations. In contrast, our method incorporates user history and reviews into the query, enabling retrieval based on richer contextual information and offering interpretability in the form of implicit rationales.

Fig. 4 presents example queries generated by BIGRec and QUEREC, along with the actual items purchased by users. While QUEREC does not explicitly mention the exact item titles, its queries align with the purchased items in terms of target user groups and product characteristics. Notably, our method generates semantically relevant queries that reflect the preferences of the user. These findings indicate that LLMs can leverage their knowledge for accurate and unbiased recommendations without requiring fine-tuning. Additional examples are included in Appendix E.3.1.

5 Conclusion

In this work, we introduced QUEREC, a novel parallel framework that enhances the diversity and performance of recommendation by using LLM to generate personalized queries. QUEREC outperforms state-of-the-art baselines including LLM-based reranking, LLM-based next-item prediction, and CF-based methods, by effectively leveraging the rich knowledge embedded in LLMs. Furthermore, our analysis highlights the ability of the method to reduce bias and increase the diversity of recommendations. By integrating the complementary strengths of LLMs and recommendation models, QUEREC provides a promising direction for future recommendation research, offering a practical, training-free solution adaptable to evolving LLMs and large-scale dynamic item pools.

6 Limitations

Despite its advantages, QUEREC has several limitations. First, while it eliminates the need for explicit candidate selection, the effectiveness of query generation depends on the quality of the pre-trained LLM and its ability to generalize across domains. In cases where the LLM lacks sufficient domain-specific knowledge, the generated queries may not capture user intent effectively. Second, since our approach relies on retrieving items based on semantic similarity, it may underperform in domains where structured interaction data (e.g., collaborative signals) are more informative than textual item representations. Finally, while QUEREC enhances diversity, it does not explicitly control for fairness, popularity bias, or serendipity, which remain open challenges for future research. Addressing these limitations through hybrid approaches, enhanced query generation mechanisms, and domain-aware enhancements presents a promising direction for further improving LLM-driven recommendation systems.

7 Ethics Statement

Potential Risks Our study was conducted on fixed datasets and the potential impact of the user in real-world applications has not been examined. Therefore, caution is required when applying the proposed method beyond the controlled experimental setting.

Use of Scientific Artifacts Our research leveraged open source tools, including PyTorch (Paszke et al., 2019), along with pre-trained language models such as LLaMA3.2 and T5 obtained via the Huggingface (Wolf et al., 2019) library. We used all artifacts in accordance with their intended purpose.

Use of Ai Assistants We only used ChatGPT to provide a better expression and to refine the language. Some of the code used in the experiment was written with the assistance of Copilot.

8 Acknowledgments

This research was supported by the National Research Foundation of Korea (RS-2025-16071337). For GPU infrastructure, our work was supported by the IITP grant funded by MSIT (No. RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale).

References

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. 2025. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27.

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*, pages 1007–1014.

Yuwei Cao, Nikhil Mehta, Xinyang Yi, Raghunandan Hulikal Keshavan, Lukasz Heldt, Lichan Hong, Ed Chi, and Maheswaran Sathiamoorthy. 2024. Aligning large language models with recommendation knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1051–1066, Mexico City, Mexico. Association for Computational Linguistics.

Aldo Carranza, Rezsa Farahani, Natalia Ponomareva, Alexey Kurakin, Matthew Jagielski, and Milad Nasr. 2024. Synthetic query generation for privacy-preserving deep retrieval systems using differentially private language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3920–3930, Mexico City, Mexico. Association for Computational Linguistics.

Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. 2024. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740*.

Zhiang Dong, Liya Hu, Jingyuan Chen, Zhihua Wang, and Fei Wu. 2025. Comprehend then predict: Prompting large language models for recommendation with semantic and collaborative data. *ACM Trans. Inf. Syst.* Just Accepted.

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*.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 299–315, New York, NY, USA. Association for Computing Machinery.

B Hidasi. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

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

Wenyue Hua, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2023. How to index item ids for recommendation foundation models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 195–204, New York, NY, USA. Association for Computing Machinery.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 1395–1406, New York, NY, USA. Association for Computing Machinery.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023a. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1258–1267.

Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023b. Gpt4rec: A generative framework for personalized recommendation and user interests interpretation. *arXiv preprint arXiv:2304.03879*.

Lei Li, Yongfeng Zhang, and Li Chen. 2023c. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Wensheng Lu, Jianxun Lian, Wei Zhang, Guanghua Li, Mingyang Zhou, Hao Liao, and Xing Xie. 2024. Aligning large language models for controllable recommendations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8159–8172, Bangkok, Thailand. Association for Computational Linguistics.

Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 825–833.

Sheshera Mysore, Andrew McCallum, and Hamed Zamani. 2023. Large language model augmented narrative driven recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 777–783.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13984, Bangkok, Thailand. Association for Computational Linguistics.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573.

Xinfeng Wang, Jin Cui, Fumiyo Fukumoto, and Yoshimi Suzuki. 2024a. Enhancing high-order interaction awareness in LLM-based recommender model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11696–11711, Miami, Florida, USA. Association for Computational Linguistics.

Xinfeng Wang, Jin Cui, Yoshimi Suzuki, and Fumiyo Fukumoto. 2024b. RDRec: Rationale distillation for LLM-based recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 65–74, Bangkok, Thailand. Association for Computational Linguistics.

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and Zhenhua Dong. 2024c. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3245–3254, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 12–22, New York, NY, USA. Association for Computing Machinery.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1611–1621.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2639–2649, New York, NY, USA. Association for Computing Machinery.

Haobo Zhang, Qiannan Zhu, and Zhicheng Dou. 2025a. Enhancing reranking for recommendation with LLMs through user preference retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 658–671, Abu Dhabi, UAE. Association for Computational Linguistics.

Haobo Zhang, Qiannan Zhu, and Zhicheng Dou. 2025b. Enhancing reranking for recommendation with LLMs through user preference retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 658–671, Abu Dhabi, UAE. Association for Computational Linguistics.

Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, Xiaofang Zhou, et al. 2019. Feature-level deeper self-attention network for sequential recommendation. In *IJCAI*, pages 4320–4326.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1893–1902.

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515.

A Clarification of contribution

In this section, we further elaborate on the aspects that distinguish our work from prior studies and clarify the key contributions of this paper.

Early research on LLM-based recommendation systems primarily relied on fine-tuning LLMs as recommendation models (Li et al., 2023a; Chen et al., 2024; Rajput et al., 2023; Geng et al., 2022). These methods either trained LLMs to generate user/item embeddings or directly predict item IDs. However, they often underutilize LLMs' reasoning abilities and incur substantial training costs due to full-parameter updates. Another recent studies have proposed injecting LLM-derived user and item information into recommendation models to leverage LLM knowledge without fine-tuning the LLMs (Xi et al., 2024; Zhang et al., 2025b). However, these methods require retraining the recommendation model to align with the LLM outputs. In contrast, we propose a training-free framework that fully leverages the pretrained LLM's reasoning capacity. This design eliminates the need for fine-tuning, reducing bias and improving generalization.

Prior work such as GPT4Rec and BIGRec (Li et al., 2023b; Bao et al., 2025) fine-tuned LLMs to generate the title-based queries. This restricts the expressiveness of the LLM and limits adaptability in dynamic item pools. Our method, by contrast, generates rich, personalized queries that incorporate user preferences and item attributes without training, enhancing retrieval quality and recommendation diversity.

Recent reranking approaches (Hou et al., 2024; Bao et al., 2023; Kim et al., 2024; Yang et al., 2023) apply LLMs to reorder a small, pre-selected candidate set. However, these candidates are typically derived from CF-based models like SASRec (Yang et al., 2023), constraining the LLM's capacity to explore diverse knowledge. We address this limitation by using LLMs to generate both user and item queries for full-pool retrieval, enabling knowledge-rich, unbiased, and adaptive recommendations. Our proposed method also avoids the positional bias caused by the "lost in the middle" phenomenon, as it does not include candidate items in the prompt.

Another line of work involves using LLMs to generate synthetic queries, which are then used to train retrievers (Mysore et al., 2023; Carranza et al., 2024). These studies have primarily been evaluated on search tasks where user intent is explicitly expressed through natural language queries. In contrast, our work focuses on recommendation settings, where user intent is implicit and not directly observable. Unlike prior efforts that use synthetic queries to train retrievers, we aim to leverage the knowledge encoded in LLMs to enrich limited user-item information and apply it to retrieval-based recommendation. Furthermore, we propose a parallel architecture that effectively integrates this augmented information with existing recommender models without requiring retraining.

B Experimental Setups

This section provides a more detailed description of the experimental setup and the baselines used.

B.1 Environment

For the query generation stage using the LLM, we employed the vLLM[2] framework, which enables efficient and scalable inference by leveraging optimized memory management and parallelization strategies.

B.2 Datasets

Dataset	#Users	#Items	#Reviews	Density (%)
Sports	35,598	18,357	296,337	0.0453
Beauty	22,363	12,101	198,502	0.0734
Toys	19,412	11,924	167,597	0.0724
Yelp	30,431	20,033	316,354	0.0519

Table 4: Statistics of the datasets.

[2]https://docs.vllm.ai/

We conducted experiments on four widely used benchmark datasets collected from the Amazon e-commerce platform: *Sports & Outdoors*, *Beauty*, *Toys & Games* and *Yelp*. Each dataset consists of user interactions, including a user ID, an item ID, a rating, a review, and a timestamp. The statistics of the data set are provided in Table 4. These datasets are widely adopted by previous studies (Li et al., 2023c; Wang et al., 2024a,b; Rajput et al., 2023; Cao et al., 2024; Geng et al., 2022), which have been extensively explored over the past three years. Due to space limitations, we include only the experimental results for the three main datasets in the manuscript, excluding the Yelp dataset. To further validate the effectiveness of our proposed method, we conducted additional experiments on the Yelp dataset, which represents a domain distinct from the other three datasets. The results of experiment on the Yelp dataset are presented in Section F.

B.3 Baselines

This subsection introduces the baselines used in our experiments and their corresponding setups. To obtain similarity scores for SASRec and DIF-SR, we trained the models using Recbole[3], a well-validated open-source recommendation framework. Detailed hyperparameters and training results will be made available through our public repository. For ELMRec, we excluded test targets affected by label leakage in the explanation task to ensure fair evaluation. In case of ELMRec, we obtain beam search scores and use them as the recommendation score. The model was trained using the same T5-small backbone and hyperparameters as in prior studies. For all other baselines, we report the performance metrics as documented in previous studies (Zhou et al., 2020; Li et al., 2023c; Wang et al., 2024a). In cases of fine-tuned LLM-based methods (e.g. TALLRec, BIGRec and GPT4Rec), we trained the LLaMA 3.2 3B with LoRA for 2 epochs with a batch size of 4 and a learning rate of 5e-5.

B.3.1 CF-base Model

- **Caser** (Tang and Wang, 2018): A CNN-based sequential recommendation model that learns item embeddings through convolution operations to effectively model users' sequential patterns.
- **GRU4Rec** (Hidasi, 2015): An RNN-based sequential recommendation model that leverages GRU to capture users' sequential patterns.
- **HGN** (Ma et al., 2019): A sequential recommendation model that captures both long-term and short-term user interests through a hierarchical gating mechanism, which selectively processes item features and instances while explicitly modeling item-item relationships.
- **SASRec** (Kang and McAuley, 2018): A sequential recommendation model that employs a self-attention mechanism to learn users' sequential behavior patterns.
- **BERT4Rec** (Sun et al., 2019): A sequential recommendation model that utilizes BERT architecture to capture users' sequential interaction patterns.
- **FDSA** (Zhang et al., 2019): A sequential recommendation model that captures both item-level and feature-level transition patterns by applying separate self-attention blocks to model their relationships, enhancing recommendation performance. It integrates heterogeneous item features using attention mechanisms and combines item and feature transitions to improve next-item prediction.
- **S^3-Rec** (Zhou et al., 2020): A self-attentive model that enhances sequential recommendation by leveraging self-supervised pre-training objectives to capture correlations among attributes, items, subsequences, and sequences using mutual information maximization.
- **DIF-SR** (Xie et al., 2022): A sequential recommendation model that enhances side information fusion by shifting it from the input to the attention layer, decoupling attention calculations to mitigate rank bottlenecks and improve gradient flexibility, thereby enhancing modeling capacity and boosting recommendation performance.

[3]https://recbole.io/

B.3.2 T5-base Model

- **P5** (Geng et al., 2022): A T5-based text-to-text recommendation model, converting all recommendation-related data into natural language sequences and utilizing personalized prompts for various recommendation tasks.

- **TIRGER** (Rajput et al., 2023): A generative model-based recommendation method, explicitly using numeric item-IDs to perform next-item prediction tasks. Unlike our approach, it does not leverage natural language generation, focusing instead on digit-based input-output structures.

- **POD** (Li et al., 2023c): A T5-based method, distills discrete prompts into continuous prompt vectors to better bridge user/item, aiming to improve both training efficiency and recommendation performance.

- **RDRec** (Wang et al., 2024b): A T5-based method, enhances recommendation by distilling rationales from user and item reviews, allowing a compact model to leverage preference and attribute-level explanations for improved recommendation performance.

- **ELMRec** (Wang et al., 2024a): A T5-based recommendation by enhancing whole-word embeddings to better interpret high-order user-item interactions without graph pre-training, while also addressing recency bias through a reranking mechanism to improve both direct and sequential recommendations.

B.3.3 LLM-based Reranking Method

- **LLMRank** (Hou et al., 2024): A zero-shot reranking approach using LLMs has been proposed to rerank selected candidate items. Three types of prompts were introduced: sequential, in-context learning, and recency-focused.

- **TALLRec** (Bao et al., 2023): An approach that fine-tunes LLMs using LoRA to enhance their recommendation capabilities, enabling efficient adaptation to recommendation tasks while reducing computational overhead. The initially proposed method used a binary prediction approach, whereas we modified the prompt to select the top-10 items from the candidate set used in a previous study (Kim et al., 2024). During training, the prompt included the top-20 candidate items recommended by the traditional recommendation model, while the training target was a reordered list of 10 item titles, ensuring that the label item appeared first. This setup was designed to allow a fair comparison that closely aligns with our proposed method.

- **A-LLMRec** (Kim et al., 2024): An LLM-based framework that integrates pre-trained CF-based user/item embeddings into LLMs, enabling effective recommendations in both cold and warm scenarios. It leverages both collaborative signals and LLM reasoning capabilities by introducing a alignment layer that requires training to bridge the two representations.

B.3.4 LLM-based Retrieval Method

- **GPT4Rec** (Li et al., 2023b): An LLM-based recommendation method that fine-tunes language models to predict a user's next item and uses the generated item text as search queries.

- **BIGRec** (Bao et al., 2025): A bi-step grounding method for LLM-based recommendation systems. It first aligns the LLM with the recommendation task by training it to generate item titles. These generated titles are then used as textual queries to retrieve items, thereby grounding the language model outputs to the recommendation space.

C Experiments on Various LLMs

To evaluate the generalizability of our proposed approach across different LLM architectures, we conducted experiments using various models, including LLaMA 3.1 8B, LLaMA 3.3 70B, Gemma 2 2B, and Gemma

2 3B. Table 5 presents the performance of QUEREC in a standalone setting. For the 70B model, we utilized *Llama-3.3-70B-Instruct-Turbo* in DeepInfra[4] for generation.

The results demonstrate that our method maintains stable performance across different LLMs, confirming its adaptability to various model architectures. Additionally, we observe that even lightweight models can achieve performance comparable to larger models, highlighting the efficiency and scalability of our approach.

Model	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
LLaMA 3.2 3B	0.0280	0.0188	0.0403	0.0228	0.0501	0.0354	0.0675	0.0410	0.0680	0.0472	0.0937	0.0555
LLaMA 3.1 8B	0.0270	0.0185	0.0397	0.0225	0.0491	0.0346	0.0667	0.0402	0.0665	0.0466	0.0926	0.0550
LLaMA 3.3 70B	0.0269	0.0182	0.0399	0.0224	0.0505	0.0361	0.0686	0.0419	0.0678	0.0469	0.0935	0.0552
Gemma 2 2B	0.0272	0.0186	0.0395	0.0225	0.0496	0.0356	0.0682	0.0416	0.0661	0.0457	0.0927	0.0543
Gemma 2 9B	0.0247	0.0165	0.0390	0.0211	0.0481	0.0345	0.0680	0.0409	0.0664	0.0460	0.0940	0.0549

Table 5: Performance variation of QUEREC across different LLMs. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

To investigate whether previously proposed reranking methods become more effective when utilizing a larger model, we conducted experiments using the LLaMA 3.3 70B. Table 6 presents the results of these experiments. Compared to lightweight models, we observed an overall improvement in performance compare to the cases used LLaMA 3.2 3B, and in some metrics, performance improvements were also observed compared to the standalone recommendation model. However, in most cases, performance degradation was still evident, reaffirming that the reranking approach remains less effective than our proposed method. These findings highlight that simply scaling up the model does not necessarily resolve the limitations of existing reranking strategies.

Model	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
DIF-SR	<u>0.0360</u>	0.0197	<u>0.0542</u>	0.0256	<u>0.0576</u>	0.0336	<u>0.0901</u>	<u>0.0442</u>	0.0700	0.0404	<u>0.0995</u>	0.0499
+ LLMRank ICL	0.0315	0.0215	0.0409	0.0246	0.0453	0.0322	0.0501	0.0339	0.0663	0.0478	0.0843	0.0537
+ LLMRank Seq	0.0348	0.0246	0.0431	0.0274	0.0489	0.0361	0.0534	0.0376	0.0725	0.0546	0.0891	0.0601
+ LLMRank Recent	0.0354	<u>0.0252</u>	0.0435	<u>0.0279</u>	0.0491	<u>0.0364</u>	0.0540	0.0380	<u>0.0739</u>	<u>0.0556</u>	0.0904	<u>0.0610</u>
QUEREC (+ DIF-SR)	**0.0389**	**0.0243**	**0.0604**	**0.0312**	**0.0715**	**0.0479**	**0.1059**	**0.0589**	**0.0868**	**0.0618**	**0.1231**	**0.0735**

Table 6: Experimental results on LLaMA 3.3 70B. The best and second-best results for each metric are highlighted in bold and underlined, respectively. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

D Divergent Perspective Reranking

In this section, we present the analysis and experiments that motivate our proposed Divergent Perspective Reranking approach. Through these studies, we demonstrate that our method offers greater stability and robustness compared to traditional ensemble techniques such as convex combination (CC) or reciprocal rank fusion (RRF).

D.1 Intersection Between Recommendation Methods

In this subsection, we present an analysis that motivated our proposed adaptive weighting scheme. Specifically, we examine the overlap in the recommended items between different recommendation methods, based on whether the recommended items are hits under the Hit@10 metric. Fig.6, 7, and 8 illustrate how the degree of intersection varies across methods and datasets. When the intersection is high, assigning weights based on the accuracy of each model can help improve the overall recommendation performance. However, when the intersection is low, skewed weights may lead to scenarios where incorrect recommendations from one model disproportionately influence the final result. To address this

[4]https://deepinfra.com/

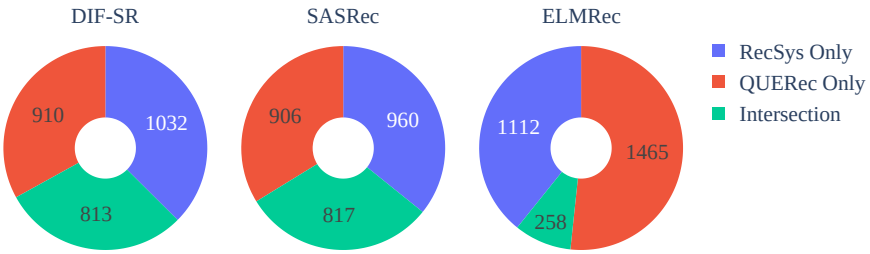


Figure 6: Intersection between recommendation methods in Toys dataset.

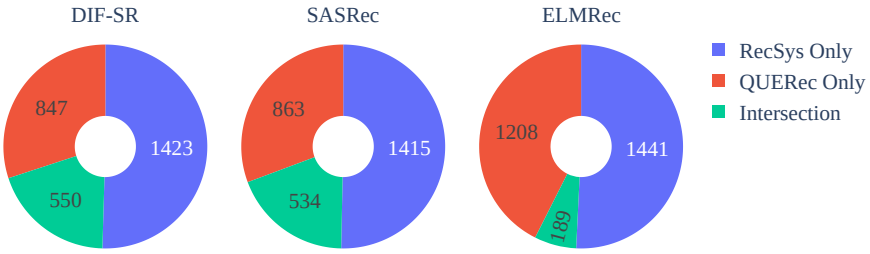


Figure 7: Intersection between recommendation methods in Beauty dataset.

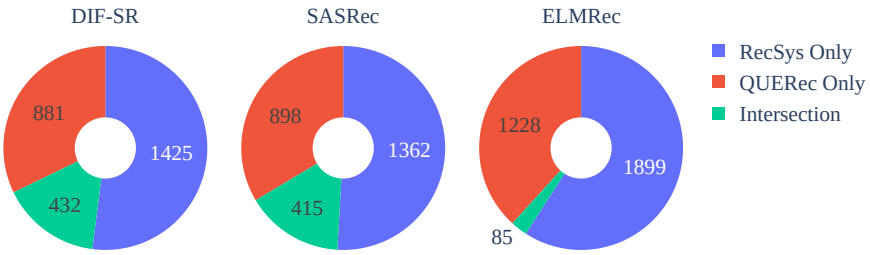


Figure 8: Intersection between recommendation methods in Sports dataset.

issue, we observe that using uniform weights (e.g., 0.5) in low-intersection cases can better optimize the final performance. Based on this insight, we propose an adaptive weighting strategy that considers not only the individual performance of each model but also the degree of intersection between their recommended items.

To integrate the top-k selection results from both the traditional recommendation model and the LLM-based approach, we employed a linear combination method. To prevent performance imbalance between the two models from negatively affecting the final ranking, we determined the λ value based on validation Hit@10 scores and intersection ratios. To evaluate the effectiveness of this approach, we compared the performance of our method against two alternatives: fixing the λ value and employing Reciprocal Rank Fusion (RRF). Table 7 presents the results of these experiments.

The results indicate that our proposed method achieved the highest performance in most cases. However, in the case of ELMRec, the RRF-based approach demonstrated superior results. This suggests that text-to-text-based recommendation models generate distinct ranking patterns compared to traditional recommendation models, highlighting the need for further research on hybrid recommendation methods that dynamically incorporate multiple recommendation systems. Nonetheless, our proposed method remains a practical solution that can be stably applied across various models with minimal overhead.

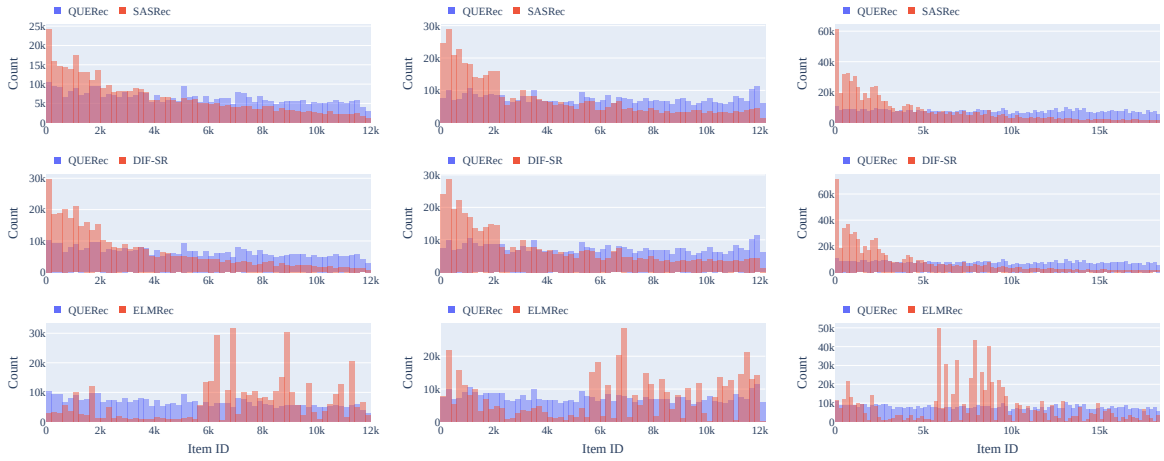
Model	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
SASRec	0.0328	0.0179	0.0499	0.0234	0.0561	0.0323	0.0872	0.0423	0.0628	0.0354	0.0915	0.0447
+ QUEREC CC (Ours)	0.0367	**0.0242**	0.0571	**0.0307**	**0.0675**	**0.0462**	**0.1022**	**0.0574**	**0.0856**	**0.0594**	**0.1189**	**0.0701**
+ QUEREC CC (0.9)	0.0272	0.0179	0.0388	0.0216	0.0488	0.0340	0.0673	0.0400	0.0681	0.0458	0.0949	0.0545
+ QUEREC CC (0.7)	0.0317	0.0209	0.0452	0.0253	0.0565	0.0391	0.0816	0.0472	0.0796	0.0538	0.1089	0.0632
+ QUEREC CC (0.5)	0.0347	0.0231	0.0534	0.0291	0.0638	0.0434	0.0988	0.0547	0.0837	0.0588	0.1180	0.0699
+ QUEREC CC (0.3)	**0.0371**	0.0219	**0.0576**	0.0285	0.0655	0.0421	0.0988	0.0529	0.0774	0.0521	0.1107	0.0628
+ QUEREC CC (0.1)	0.0345	0.0192	0.0522	0.0249	0.0597	0.0360	0.0907	0.0460	0.0675	0.0401	0.0978	0.0498
+ QUEREC RRF	0.0330	0.0198	0.0526	0.0261	0.0585	0.0374	0.0899	0.0475	0.0730	0.0473	0.1056	0.0578
DIF-SR	0.0360	0.0197	0.0542	0.0256	0.0576	0.0336	0.0901	0.0442	0.0700	0.0404	0.0995	0.0499
+ QUEREC CC (Ours)	0.0371	**0.0247**	**0.0590**	**0.0318**	**0.0699**	**0.0472**	**0.1032**	**0.0579**	**0.0885**	**0.0615**	**0.1228**	**0.0726**
+ QUEREC CC (0.9)	0.0274	0.0180	0.0389	0.0217	0.0489	0.0340	0.0676	0.0401	0.0690	0.0461	0.0949	0.0545
+ QUEREC CC (0.7)	0.0327	0.0214	0.0465	0.0259	0.0580	0.0402	0.0833	0.0482	0.0804	0.0541	0.1098	0.0635
+ QUEREC CC (0.5)	0.0360	0.0239	0.0557	0.0303	0.0668	0.0457	0.0982	0.0558	0.0842	0.0594	0.1204	0.0711
+ QUEREC CC (0.3)	**0.0380**	0.0227	0.0580	0.0292	0.0654	0.0437	0.1003	0.0550	0.0770	0.0537	0.1120	0.0649
+ QUEREC CC (0.1)	0.0346	0.0194	0.0542	0.0257	0.0622	0.0381	0.0910	0.0473	0.0680	0.0427	0.0998	0.0530
+ QUEREC RRF	0.0338	0.0202	0.0532	0.0264	0.0608	0.0390	0.0920	0.0491	0.0731	0.0483	0.1073	0.0593
ELMRec	0.0492	0.0414	0.0569	0.0437	0.0610	0.0503	0.0729	0.0540	0.0706	0.0616	0.0749	0.0623
+ QUEREC CC (Ours)	**0.0589**	0.0455	**0.0746**	**0.0506**	**0.0816**	**0.0625**	**0.1060**	**0.0704**	**0.1092**	0.0777	**0.1350**	0.0860
+ QUEREC CC (0.9)	0.0256	0.0170	0.0381	0.0211	0.0477	0.0331	0.0648	0.0386	0.0658	0.0445	0.0919	0.0529
+ QUEREC CC (0.7)	0.0337	0.0217	0.0532	0.0279	0.0592	0.0401	0.0867	0.0489	0.0831	0.0564	0.1172	0.0674
+ QUEREC CC (0.5)	0.0586	**0.0462**	0.0717	0.0504	0.0805	0.0617	0.1026	0.0689	0.1032	**0.0811**	0.1267	**0.0887**
+ QUEREC CC (0.3)	0.0504	0.0427	0.0594	0.0456	0.0665	0.0545	0.0821	0.0595	0.0786	0.0676	0.0936	0.0724
+ QUEREC CC (0.1)	0.0491	0.0444	0.0562	0.0421	0.0623	0.0515	0.0745	0.0555	0.0681	0.0621	0.0721	0.0634
+ RRF	0.0576	0.0447	0.0720	0.0494	0.0791	0.0595	0.1014	0.0667	0.0957	0.0752	0.1208	0.0833

Table 7: Experimental results based on various CF-LLM cooperation reranking methods. The best and second-best results for each metric are highlighted in bold and underlined, respectively. (Ours) denotes our proposed Hit Rate-based reranking method, while the other numerical values represent fixed lambda (λ) values used in the linear combination approach. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

E Additional Analysis

E.1 Item Distribution

We stated that our proposed method has the potential to mitigate bias introduced by the training dataset. Fig. 9 illustrates that major baselines exhibit strong bias toward certain items, whereas our method provides more balanced recommendations across a wide range of items. To investigate whether this bias originates from the training data, we conducted an additional analysis to examine the distribution of item interactions in the training set.



(a) Toys. (b) Beauty. (c) Sports.

Figure 9: Diversity of recommendations across different datasets and models. The x-axis represents the IDs of the recommended items, while the y-axis indicates the frequency of recommendation. The top-20 recommended items for each user were extracted and compared.

(a) Toys. (b) Beauty. (c) Sports.

Figure 10: Target item ID distribution across different datasets. The x-axis represents the IDs of the target items, while the y-axis indicates the frequency of items.

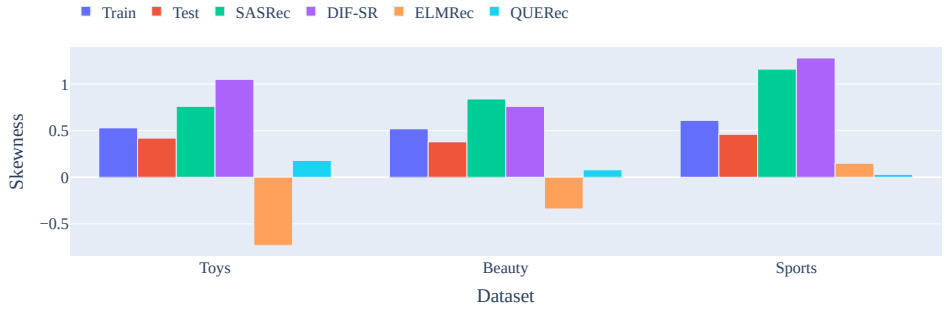


Figure 11: Comparison of item distribution skewness across datasets and the skewness of recommendation results for each model.

We conducted a more detailed analysis of the bias present in conventional recommendation systems. Fig. 10 compares the distribution of item IDs in the training and test datasets used in our experiments. While the overall distributions are similar, we observe that the training dataset is more skewed towards lower item IDs, whereas the test dataset contains a relatively higher frequency of items with larger IDs.

To quantify this observation, we measured the *skewness* of the distributions. Fig. 11 presents a comparison of the skewness values for both datasets and the recommendation outputs of various models. Our analysis confirms that the training set exhibits a higher degree of skewness compared to the test set. Additionally, traditional recommendation models demonstrate an even greater level of skewness in their recommendation results than the training dataset itself.

These findings indicate that conventional recommendation approaches may contribute to reduced diversity in recommendations. This further underscores the necessity of unbiased recommendation methods, such as our proposed approach, to mitigate such biases and improve recommendation fairness.

E.2 Latency

We measured the latency of our system by generating recommendations for 100 users under our experimental setup. On average, query generation required 69.7 seconds, and dense retrieval using an encoder took 2.8 seconds, resulting in an overall latency of approximately 0.72 seconds per user. We consider this latency reasonable and acceptable for real-world recommendation systems.

While prior LLM-based methods may exhibit similar latency when handling a small candidate set, they become inefficient in realistic settings where the number of candidate items exceeds 10,000. These approaches typically include candidate items directly within the prompt, which significantly increases the input length and degrades performance during inference. To mitigate this, such methods must partition the candidate pool into smaller subsets (e.g., fewer than 20 items per prompt) and perform multiple inference passes. However, this strategy introduces latency that grows linearly with the size of the candidate pool, rendering it unsuitable for deployment in large-scale recommender systems.

In contrast, our method decouples candidate selection from LLM inference, maintaining constant latency regardless of the size of the item pool. This key property makes our approach highly efficient and scalable, offering a practical solution for real-world recommendation scenarios.

E.3 Query Quality Evaluation

In this section, we evaluate the quality of the generated queries through both quantitative and qualitative analyses.

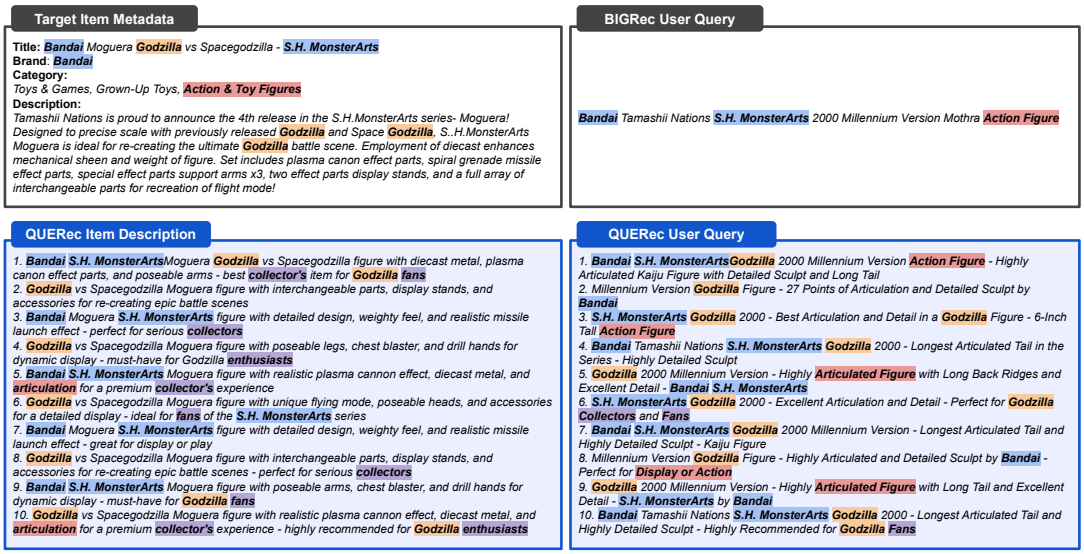


Figure 12: Query quality comparison in the Toys Dataset. Related attributes are highlighted in the same color.

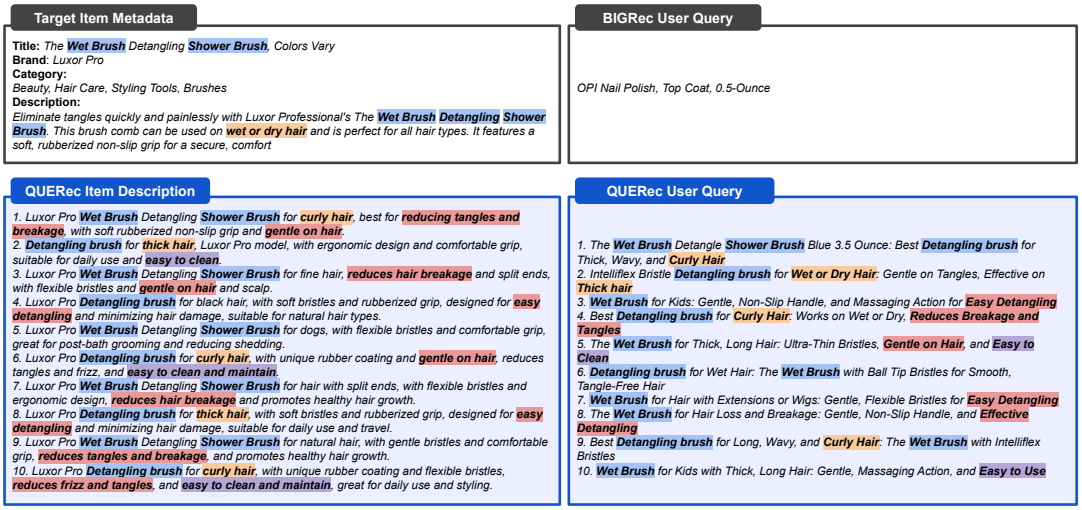


Figure 13: Query quality comparison in the Beauty Dataset. Related attributes are highlighted in the same color.

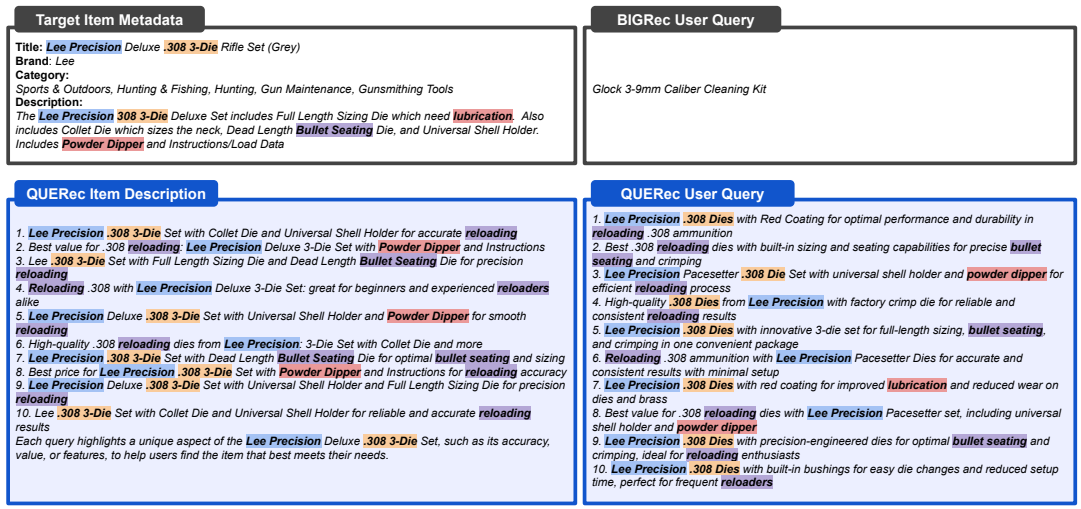


Figure 14: Query quality comparison in the Sports Dataset. Related attributes are highlighted in the same color.

E.3.1 Generated Query Comparision

Fig. 12, 13, and 14 present examples from each dataset, showcasing the personalized user queries, expanded item descriptions, and corresponding target item information. Our proposed method successfully matches key product attributes with user preferences by generating richer information without any additional training. In contrast, fine-tuned baseline approaches often align with only limited features or mention irrelevant item titles. These results highlight the effectiveness of our method in leveraging the capabilities of LLMs for recommendation.

E.3.2 Query Refinement

To examine the potential impact of query quality on recommendation performance, we conducted additional experiments involving the removal of redundant or irrelevant queries. Specifically, we utilized LLaMA 3.3 70B to filter and refine the generated queries for 3,000 users. The evaluation results are summarized in Table 8.

Query Type	Sports				Beauty				Toys			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
Original	0.0253	0.0173	0.0367	0.0209	0.0360	0.0252	0.0470	0.0287	0.0630	0.0446	0.0883	0.0528
Refined	0.0263	0.0173	0.0367	0.0205	0.0350	0.0247	0.0480	0.0289	0.0613	0.0434	0.0873	0.0518

Table 8: Comparison of original and refined queries across datasets.

Overall, the performance of refined queries were similar to the original ones, with minimal differences across most datasets. Notably, a slight performance drop was observed in the Toys domain. These findings suggest that the original queries produced by the LLM were already of high quality, providing effective signals for retrieval.

One possible explanation for the performance degradation is the loss of feature redundancy during the refinement process. In the original queries, certain attributes appeared multiple times, which may have implicitly reflected strong user preferences. By removing these repetitions, the refinement may have unintentionally weakened the representation of dominant user interests, thereby reducing the effectiveness of the final recommendations.

F Evaluation on Yelp Dataset

We utilized datasets widely adopted by previous studies (Wang et al., 2024a; Rajput et al., 2023; Li et al., 2023c; Cao et al., 2024; Wang et al., 2024b; Geng et al., 2022), which have been extensively explored in recent research over the past three years. Major publications frequently employed the three datasets we selected. To ensure reproducibility and fair comparison, we chose the same datasets, believing they adequately demonstrate the effectiveness of our proposed approach.

To further validate the generalizability of our approach, we conducted additional experiments on the *Yelp* dataset[5]. Unlike the previously used Amazon datasets, which are primarily focused on e-commerce domains (e.g., Beauty, Sports, Toys), Yelp consists of user-generated reviews centered around local businesses such as restaurants and services, presenting a distinct domain with different user behavior patterns and item characteristics. This expansion allows us to examine the adaptability of our method beyond typical product recommendation settings. Yelp has been widely adopted in recent recommendation studies (Dong et al., 2025; Zhou et al., 2020; Wang et al., 2024c; Hua et al., 2023), making it a reliable benchmark for evaluating model effectiveness across domains.

Our method continues to demonstrate competitive or superior performance on this dataset, further supporting its applicability in diverse real-world scenarios. These results confirm that the advantages of our training-free, modular architecture—such as performance gains and improved diversity—are not confined to a specific dataset or domain, but extend to varied recommendation environments.

[5]https://www.yelp.com/dataset

Model	Yelp			
	H@5	N@5	H@10	N@10
Caser	0.0151	0.0096	0.0253	0.0129
GRU4Rec	0.0152	0.0099	0.0263	0.0134
HGN	0.0186	0.0115	0.0326	0.0159
SASRec	0.0452	0.0334	0.0630	0.0391
BERT4Rec	0.0051	0.0033	0.0090	0.0045
FDSA	0.0271	0.0170	0.0464	0.0232
S^3-Rec	0.0168	0.0123	0.0341	0.0168
DIF-SR	0.0452	0.0335	0.0651	0.0398
P5	0.0225	0.0159	0.0329	0.0193
TIGER	0.0212	0.0146	0.0367	0.0194
QUEREC (+ SASRec)	<u>0.0506</u>	<u>0.0408</u>	<u>0.0684</u>	<u>0.0465</u>
QUEREC (+ DIF-SR)	**0.0511**	**0.0410**	**0.0700**	**0.0471**
QUEREC (w/o CF-based model)	0.0314	0.0267	0.0376	0.0287

Table 9: Performance comparison on the Yelp dataset. The best and second-best results for each metric are highlighted in bold and underlined, respectively. "H@k" and "N@k" denote Hit Rate and NDCG at rank k, respectively.

G Prompt and Generated Query

This section presents the prompts used in our query generation approach along with examples of the generated queries. We designed the prompt format based on the LLaMA3 instruction format. For user history, we set the maximum length to 8 and included a phrase to emphasize the last interaction. To prevent excessive prompt length, we incorporated item queries only for the last interaction during user query generation. We instructed the LLMs to generate 10 queries per each item and user. We utilized vLLM [6] to perform the generation steps required for training-free methods, including query generation. The complete set of generated queries will be released through an online public repository upon acceptance.

G.1 Item Query Generation Prompt

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an intelligent assistant designed to create detailed and precise search queries for items based on their descriptions and
      aggregated user reviews.
Your task is to generate 10 distinct and comprehensive search queries that effectively help users find the specified item.
Focus on incorporating key features, standout aspects, brand, and practical benefits into each query to enhance search accuracy.
Emphasize the unique attributes that differentiate the item from similar products.
Each query should be concise, factual, and separated by line breaks.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
### Task:
Analyze the provided item metadata and user reviews to generate 10 detailed and objective search queries for the item.
Your goal is to create queries that highlight the item's key features, benefits, and unique aspects based on its description and
      user feedback.

### Input:
- **Item Title**: Last Night on Earth: Growing Hunger Expansion
- **Brand**: Flying Frog Productions
- **Description**: The Growing Hunger Expansion introduces new game mechanics and three exciting new Scenarios to challenge
      players as well as a two player mini-game. Take control of four new Heroes, each with a highly-detailed plastic miniature as
       well as seven new Red Zombies for use as Plague Carriers, Grave Dead, or to increase the Zombie Horde. New modular game
      board sections expand the town and feature unique buildings such as the Supermarket, Library, and Antique Shop. New game
      cards give Zombies a chance to steal weapons from the Heroes and add powerful Double-Handed weapons to the Heroes arsenal,
      such as Garden Shears and the Fence Post. Also included are two new full color, die-cut counter sheets adding Free Search
      Markers for the Heroes as well as many more fun ideas to the Last Night on Earth toolbox for limitless use with official web
       content or creating your own new Scenarios.
-**User Reviews**:
1. Four Stars Great games
2. Last Night on Earth is the best! Last Night on Earth: Growing Hunger is an expansion for the Last Night on Earth game. I love
      it because it is so much fun! My friends and I get together once a month for game night and this is one we always pull out.
      This expansion add heroes, props, locations and scenarios to the main game. You do need the main Last Night on Earth board
      game for this add on to help you in any way. If you would like to know more about the main game there is a video on you tube
       from Wil Wheaton on his tabletop gaming blog style show called "TableTop" they played last night on earth https://www.
      youtube.com/watch?v=UhLU2-BuhMIIts a great explanation of the game and how it works! However I took it to the next level and
       I hand painted my hero figures. I will post the images.
3. Good expansion It's a good expansion for a good game. I don't know what else to write so I'm finishing with filler.
4. Great expansion to a great game A worthy expansion to the original game, adding new characters, cards, scenarios, and lots and
      lots of extra components for building your own scenarios and adventures. For me, thats the big difference between owning
      just the game and this expansion: whether or not you intend to make your own scenarios or just play the ones that come with
      the game. If you are one of those "game tinkerers" like me, then you will love this expansion.
```

[6] https://github.com/vllm-project/vllm

```
5. A good addition to a great game. I'll keep it pretty short. If you're considering the expansion, I would hope you already own
      Last Night on Earth. I think LNOE is a great game that offers suspense, teamwork, competition, and usually a lot of laughs.
      This expansion does not change much. What it does do is add some more locations (in the form of some new L-shaped boards),
      some new characters, new hero and zombie cards, and a few new scenarios. On top of that, it comes with new game pieces that
      are used with the free scenarios available on the games' official website or that can be used to design your own scenarios.
      In short, this expansion does not reinvent the wheel. What it will do, however, is help expand and reinvigorate an already
      great game. It's a little pricey, but if you like LNOE, it's a good investment.
6. great game add on The is a great game expansion to the game last night on earth, its very fun and different i highly recommend
      this add on. love the new game pieces


### Requirements:
- Generate exactly 10 distinct search queries, each on a **separate line**.
- Incorporate key metadata and review insights to create effective, descriptive queries.
- Highlight the item's purpose, standout features, brand, and practical benefits in each query.
- Emphasize the **unique attributes that differentiate the item from other similar products**.
- Avoid redundant details and ensure each query is unique and precise.

### Response:
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

G.2 Item Query Sample

Toys

```
1. **"Radica 20Q AI Game for Kids and Adults: Color Options and Educational Value"**
2. **"20Q Toy with Smart Features: How Does it Guess What You're Thinking?"**
3. **"Best Party Game for Adults and Kids: 20Q Review and Comparison"**
4. **"Portable and Durable 20Q Game for Travel and Classroom Use"**
5. **"Radica 20Q AI Game: How Accurate is it in Guessing Words and Concepts?"**
6. **"Fun and Educational Toy for Kids: 20Q Review and Gift Ideas"**
7. **"20Q Game for Large Groups: Conversation Starter and Icebreaker"**
8. **"Best Gift for Kids and Adults: 20Q Artificial Intelligence Game with Reviews"**
9. **"20Q Toy with Interactive Features: How Does it Engage Kids and Adults?"**
10. **"Radica 20Q Game: Unique Features and Benefits for Home and Classroom Use"**
```

Beauty

```
1. **"Fashionwu 10 Pair Long Black False Eyelashes Makeup for Everyday Use, Natural Look with 1cm-1.5cm Length and 3cm Width"**
2. **"Long Black False Eyelashes for Beginners, 10 Pairs with Curved Shape, Perfect for Enhancing Natural Lashes"**
3. **"Fashionwu Black False Eyelashes for Cosplay, 10 Pairs with Thick and Sturdy Design, Great for Drag Queens and Performers"**
4. **"Affordable 10 Pair Long Black False Eyelashes, Good for Special Occasions and Everyday Wear, Natural and Sassy Look"**
5. **"Fashionwu Black False Eyelashes with Glue Strip, 10 Pairs for Long-Lasting Wear, Perfect for Lash Enthusiasts"**
6. **"Long Black False Eyelashes for Beginners, 10 Pairs with Thicker Band and Longer Length, Great for Volume and Drama"**
7. **"Fashionwu Black False Eyelashes for Stage Performers, 10 Pairs with Sturdy Design and Long-Lasting Adhesive"**
8. **"Natural and Sassy Black False Eyelashes, 10 Pairs with Curved Shape and Thicker Band, Perfect for Everyday Wear"**
9. **"Fashionwu Black False Eyelashes for Special Occasions, 10 Pairs with Glitter and Sparkle, Great for Cosplay and Parties"**
10. **"Long Black False Eyelashes for Lash Enthusiasts, 10 Pairs with Thicker Band and Longer Length, Perfect for Volume and Drama
      "**
```

Sports

```
1. **Trijicon NS 3Dot Set (GR/GR) T0963 for Glock: Military-Tested, Bright & Tough Night Sights with 12-Year Tritium Life**
2. **Best Trijicon Night Sights for Glock: 3Dot Set with White Rings for Enhanced Daylight Visibility and Shock Protection**
3. **Trijicon NS 3Dot Set (GR/GR) T0963: Proven Reliability and Accuracy for Combat Handguns**
4. **Trijicon Tritium Night Sights for Glock: 3Dot Set with Aluminum Housing and Silicon Rubber Cushions for Low-Light Performance
      **
5. **Trijicon NS 3Dot Set (GR/GR) T0963: 12-Year Warranty, Military-Grade Night Sights for Glock and Other Handguns**
6. **Trijicon NS 3Dot Set (GR/GR) T0963: Fast Target Acquisition and Superior Daylight Visibility with White Rings**
7. **Trijicon NS 3Dot Set (GR/GR) T0963 for Glock: Shock-Resistant and Durable Tritium Night Sights**
8. **Trijicon Tritium Night Sights for Glock: 3Dot Set with Enhanced Low-Light Performance and Long-Lasting Tritium Life**
9. **Trijicon NS 3Dot Set (GR/GR) T0963: Fast and Accurate Night Sights for Glock Handguns with White Rings and Aluminum Housing**
10. **Trijicon NS 3Dot Set (GR/GR) T0963: Military-Tested, Proven Night Sights for Glock and Other Handguns with 12-Year Warranty
      **
```

G.3 User Query Generation Prompt

G.3.1 Prompt Format

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an intelligent assistant designed to analyze a user's purchase history and behavior to generate **10 effective search
      queries** for predicting the **next items** they are most likely to purchase.
Your task is to evaluate past purchase patterns, item metadata, and related search queries to construct concise and accurate
      search queries that can be used to find the next recommended items.
Focus on identifying recurring patterns, shifts in preferences, and evolving interests to enhance the relevance of the search
      queries.
For the most recent item, make sure to include **related queries** that were associated with it to improve search accuracy.
Ensure each query highlights the **unique characteristics of items** and reflects the **user's preferences and interests** for
      more personalized recommendations.
Ensure each query is clear, specific, and optimized for retrieving relevant items.
```

```
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
### Task:
You are an intelligent assistant tasked with generating **10 optimized search queries** to predict the **next items** a user is
      likely to purchase based on their chronological purchase history, item metadata, and related search queries.

**Purchase History:** A chronological list of items the user has purchased, including item brands, categories, descriptions,
      associated metadata, and related search queries. For the **most recent item**, related queries are also provided to enhance
      search relevance.


### Output Format:
Your response should follow this exact format, ensuring:
1. Each search query is presented on a **separate line**.
2. **No newlines or additional formatting** within each query.
3. The queries should be concise, specific, and optimized for accurate item retrieval.


### Requirements:
- Generate **10 precise search queries** based on the user's purchase history, item metadata, and related search queries.
- For the **most recent item**, ensure that **related queries** are incorporated to improve relevance.
- Ensure each query captures key patterns, preferences, and interests derived from the provided data.
- **Highlight the unique characteristics of items** (e.g., special features, distinctive attributes) and reflect the **user's
      preferences and behavioral trends** in the queries.
- Do **not** include explanations, introductions, or follow-up comments.
- Keep each query **clear, concise, and limited to a single line**.


### Input:
{user_history}
{generated_queries}

### Output:<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

G.3.2 User History Example

Toys

```
**Title:** `Melissa &amp;amp; Doug Wooden Take Along Tool Kit (24pc)`
**Brand:** Melissa &amp;amp; Doug
**Categories:** Toys & Games, Dress Up & Pretend Play, Pretend Play, Construction Tools
**User Review:**
Daughter loves to be like daddy My little girl loves to play pretend and "help" her Daddy and Poppa. She loves to build, hammer,
      and screw nails and bolts along with them. We have had it about a year and the nails are falling apart - it's not quite as
      well built as most of the Melissa and Doug products we have had in the past. However, we are still very pleased with the
      product overall. I especially like the fine motor skills it helps to develop.

**Title:** `Melissa &amp;amp; Doug Penguin Plush`
**Brand:** Melissa &amp;amp; Doug
**Categories:** Toys & Games, Stuffed Animals & Plush, Animals & Figures
**User Review:**
One of her favorite stuffed animals My parents bought this for my daughter last Christmas, and she still loves it. It was as tall
      as she was last year, so she was adorable at Christmas hugging it and lugging it around the house. Now that she is older,
      she still loves playing with it and it is easier for her to manipulate.One reason they bought the penguin was due to her
      love of the movie "Happy Feet". One of her favorite things to do is lie on the floor with her penguin while watching the
      movie. I don't see much educational value other than supporting (or spurring) an interest in penguins.

**Title:** `Melissa &amp;amp; Doug Farm Sound Puzzle`
**Brand:** Melissa &amp;amp; Doug
**Categories:** Toys & Games, Puzzles, Pegged Puzzles
**User Review:**
Another favorite This is another great Melissa and Doug product my parents bought for my daughter. When she was two, she got this
      for her birthday. It was so cute to see her face the first time she lifted one of the pieces and heard the animal sound.
      Personally, we thought the sound quality was fine.The wooden construction provides plenty of durability, and she did not
      tear it up. We were able to pass it on to her younger cousin. The sound quality was deteriorating slightly by that time, but
       it was still relatively easy to hear. We had it for about one year before passing it to her cousin.

This is the most recently purchased product:
**Title:** `Melissa &amp;amp; Doug Flip to Win Memory Game`
**Brand:** Melissa &amp;amp; Doug
**Categories:** Toys & Games, Games, Board Games
**User Review:**
Another great product I bought this last year for my now four-year-old daughter. We have always been pleased with the Melissa and
      Doug product line, and this one does not disappoint. I thought the cardboard "sheets" that you interchange would not last,
      but they have held up surprisingly well. We only have one that is a bit dog-eared, and that's due to my daughter thinking it
       would taste delicious so into the mouth it went.It's been a great learning tool. For example, the colors let us explore
      colors far beyond the basic 6 or 8 most learning tools address. This has increased her vocabulary, and it has made learning
      much more fun.All in all, this is one of my favorite products - and my daughter likes it a ...
```

Beauty

```
**Title:** `Now Foods: Tea Tree Oil, 4 oz`
**Brand:** Now Foods
**Categories:** Beauty, Skin Care, Body, Moisturizers, Oils
```

User Review:
For keloid scars. I bought this to treat my keloid scars and well, it didn't work, but I give it a high rating anyway because it does have other benefits and uses. The smell isn't as bad, it smells like wet wood or wet bark.

Title: `Dermablend Quick Fix Concealer SPF 30, Ivory`
Brand: Dermablend
Categories: Beauty, Makeup, Face, Concealers & Neutralizers
User Review:
Works well but wrong color. The color described isn't accurate. This turned out to be a more pinkish color and I am tan/yellowish so it did not work well for me. It is really thick though so it works as far as covering blemishes.

Title: `Finulite Cellulite Smoothing Massage Mitt`
Brand: Finulite
Categories: Beauty, Bath & Body, Bathing Accessories, Bath Mitts & Cloths
User Review:
Smooth skin. I used this on my thighs with the Finulite cream and in the shower with regular body soap and it made my skin super smooth. It removes dead skin and you can see it on the massage mitt. It's easy to clean as well. It's a tad painful on dry skin but in the shower when your skin is wet, it glides very smoothly and feels really nice.

Title: `Xtreme Brite Brightening Gel 1oz.`
Brand: Xtreme Brite
Categories: Beauty, Hair Care, Styling Products, Creams, Gels & Lotions
User Review:
Mixed feelings. I have mixed feelings about this product. When I first started using it, I can see and feel the difference. It actually started to literally peel the dark skin away. There's a bit of a sting but that's how I know it was working. My skin was lighter in just a couple weeks! However, once I've stopped using it when I reached my desired skin color, it went back to being dark. I applied more and suddenly it was no longer working. It didn't peel or sting. I wonder what went wrong or if it has an expiration date or something. Overall, it works but definitely not consistent.

Title: `Remington CI95AC/2 Tstudio Salon Collection Pearl Digital Ceramic Curling Wand, 1/2 Inch - 1 Inch`
Brand: Remington
Categories: Beauty, Hair Care, Styling Tools, Irons, Curling Irons
User Review:
My favorite tool! These are great for natural, beachy waves for your hair. The heatproof gloves work wonders as well. It heats up real hot and is super easy to use.

This is the most recently purchased product:
Title: `Finulite - The End to Cellulite, AM/PM Cellulite Cream (2 - 4 oz tubes)`
Brand: Unknown
Categories: Beauty, Skin Care, Body, Moisturizers, Creams
User Review:
I guess it's my fault. The main thing with these creams is you have to be diligent and consistent, which I wasn't. It is a PAIN in the butt to do it every night and every morning, scrubbing with the somewhat painful glove I bought with it. It did make my skin super smooth, but cellulite is a bitch to get rid of even with exercise (lots of thin people I know have cellulite, so it's definitely not a fat people thing). Long story short, I got tired of the routine and gave up.

Sports

Title: `Tone Fitness Cement Filled Kettlebell Set - 30 lbs.`
Brand: Tone Fitness
Categories: Sports & Outdoors, Exercise & Fitness, Strength Training Equipment, Kettlebells
User Review:
EXCELLENT Prouct EXCELLENT kettlebell's for both men and women. Vinyl is very smooth and easy on the hands. I would highly recommend.

Title: `Blackburn AirTower 2 Bicycle Pump, Silver`
Brand: Unknown
Categories: Sports & Outdoors, Cycling, Accessories, Bike Pumps, Floor Pumps
User Review:
Dy-no-mite This think can fill an auto tire in minutes. The best pump I have ever had. Highly recommend it. Not cheap thin metal. This is good quality thick steel.

Title: `Kimber Pepperblaster 2 Red, One Size`
Brand: Kimber
Categories: Sports & Outdoors, Outdoor Gear, Camping & Hiking, Personal Care
User Review:
Welll now... This shoots two shots of Pepper spray. That is it. You throw it away. It is NOT reloadable. It is the same excellent Kimber quality as their weapons are... But how do I know where to aim if I can't practice shoot it ??? I would not recommend this to anyone.

Title: `Ultimate Arms Gear Tactical 4 Reticle Red Dot Open Reflex Sight with Weaver-Picatinny Rail Mount`
Brand: Ultimate Arms Gear
Categories: Sports & Outdoors, Hunting & Fishing, Hunting, Hunting Optics, Gun Scopes, Rifle Scopes
User Review:
Excellent Fits my gun great. Easy to mount. Very large so the image is clear. No reason you should ever miss a target with this site. GREAT quality.

Title: `Camelbak Podium Big Chill 25 oz Bottle`
Brand: CamelBak
Categories: Sports & Outdoors, Accessories, Sports Water Bottles
User Review:
Ahhh.... Hard to get enough water out of this bottle. Doesn't keep anything cold. Very thin material. I am worried after being out in the sun, how long before it breaks.

Title: `Kimber Pepperblaster Ii Holster`
Brand: Meprolight

```
**Categories:** Sports & Outdoors, Paintball & Airsoft, Airsoft, Holsters
**User Review:**
As I said in the other feedback If you think you need the Kimber Pepperblaster then this is a great holster to carry it in. Now
      that I found out you can't reload you Pepperblaster, I see no reason for either the blaster or the holster. How do you
      practice ...???

**Title:** `5 LED Bicycle Rear Tail Red Bike Torch Laser Beam Lamp Light`
**Brand:**
**Categories:** Sports & Outdoors, Cycling, Lights & Reflectors, Taillights
**User Review:**
Great lite Very bright. I would recommend this lite to anyone. I put one on my wife's bike. I love this light.

This is the most recently purchased product:
**Title:** `Body Solid MA307N Nylon Head Harness`
**Brand:** Body Solid
**Categories:** Sports & Outdoors, Exercise & Fitness, Exercise Machine Accessories, Exercise Machine Attachments
**User Review:**
Exactly what I wanted My friend runs a boxing school for urban kids. These kids work hard at staying out of trouble, doing their
      school work and being able to learn to box. This harness works perfect. It is of highest quality.
```

G.4 User Query Sample

G.4.1 Toys

```
1. "LeapFrog TAG Reading System Purple 32MB pen with phonetic sound and games for 4-year-olds, educational toy for learning to
      read and spell"
2. "Best LeapFrog reading system for preschoolers with built-in games and activities, suitable for travel and home use"
3. "LeapFrog TAG Reading System Purple review, educational toy for kids to learn alphabet, spelling, and reading skills with
      phonetic sound and games"
4. "LeapFrog TAG Reading System 32MB pen with LeapFrog Connect software for downloading and managing content, suitable for
      children aged 3-6 years"
5. "LeapFrog TAG Reading System Purple with headphones, compact and portable learning tool for kids to learn to read and spell
      with fun games and activities"
6. "LeapFrog TAG Reading System Purple review, educational toy for kids to learn phonics, sight words, and reading skills with
      LeapFrog Connect software and games"
7. "LeapFrog TAG Reading System Purple 32MB pen with built-in games and activities, suitable for kids to learn to read and spell
      with phonetic sound and fun stories"
8. "LeapFrog TAG Reading System Purple with LeapFrog Connect software for downloading and managing content, suitable for children
      aged 3-6 years, educational toy for learning to read and spell"
9. "LeapFrog TAG Reading System Purple review, educational toy for kids to learn alphabet, spelling, and reading skills with
      phonetic sound, games, and activities"
10. "LeapFrog TAG Reading System Purple 32MB pen with LeapFrog Connect software and games, suitable for kids to learn to read and
      spell with fun stories and activities, compact and portable"
```

G.4.2 Beauty

```
1. "Neutrogena Microdermabrasion System with glycerin and ultra-fine crystals for smooth, luminous skin and anti-aging benefits,
      clinically proven for visible results in one use."
2. "Best affordable microdermabrasion system for acne-prone skin, gentle exfoliation, and firming, with 12 pre-dosed puffs and AA
      batteries included."
3. "Neutrogena Microdermabrasion System with massaging micro-vibrations for improved skin texture, radiance, and fine line
      reduction, suitable for sensitive skin types."
4. "Microdermabrasion system for at-home use, with dermatologist-recommended Neutrogena brand, proven to deliver smoother, more
      luminous skin in just one use, and affordable price point."
5. "Exfoliating microdermabrasion system with gentle, pre-dosed puffs and soothing glycerin, ideal for daily use, with visible
      results in just one treatment, and suitable for all skin types."
6. "Neutrogena Microdermabrasion System with anti-aging benefits, firming, and skin brightening, with a unique combination of
      exfoliating crystals and micro-vibrations, and easy to use at home."
7. "Best microdermabrasion system for oily skin, with a gentle, non-irritating formula, and visible results in just one use, with
      a affordable price and convenient packaging."
8. "Neutrogena Microdermabrasion System with a dermatologist-recommended formula, proven to deliver smoother, more radiant skin,
      and suitable for daily use, with a unique combination of exfoliation and micro-vibrations."
9. "Microdermabrasion system for anti-aging, firming, and skin brightening, with a gentle, pre-dosed puff system, and visible
      results in just one treatment, and suitable for sensitive skin types."
10. "Neutrogena Microdermabrasion System with a clinically proven formula, delivering visible results in just one use, and a
      affordable price point, with a unique combination of exfoliation and micro-vibrations for smoother, more luminous skin."
```

G.4.3 Sports

```
1. "ProSource Heavy-Duty Easy Gym Doorway Chin-Up/Pull-Up Bar with 300lb weight capacity and multi-position design for home
      workout"
2. "Best doorway pull-up bar for thin walls and easy installation with ProSource Comfort Grip technology"
3. "Heavy-duty pull-up bar for home gym with adjustable grip and sturdy construction for 24-32 inch doorways"
4. "ProSource Easy Gym Doorway Chin-Up/Pull-Up Bar with wall-mounting option and 5-star durability rating"
5. "Inexpensive and easy-to-assemble pull-up bar for home workout with 300lb weight capacity and multi-functional design"
6. "Best pull-up bar for doorways with raised molding and minimal wall damage with ProSource brand guarantee"
7. "Heavy-duty doorway pull-up bar with adjustable grip and comfortable grip technology for home gym workouts"
8. "ProSource Heavy-Duty Easy Gym Doorway Chin-Up/Pull-Up Bar with 300lb weight capacity and easy installation for home fitness"
9. "Best pull-up bar for doorways with multi-position design and sturdy construction for home workout and exercise"
10. "ProSource Easy Gym Doorway Chin-Up/Pull-Up Bar with 300lb weight capacity and wall-mounting option for home gym and fitness
      enthusiasts"
```