ReList: A Multi-objective Reasoning Framework for Diversified Listwise Query Recommendation

Shuxian Bi[1]*, **Chenxu Wang**[2], **Wenjie Wang**[1]†, **Yueqi Mou**[2],
Fuli Feng [1]† **Tang Biao**[2], **Peng Yan**[2],
[1]University of Science and Technology of China, [2]Meituan
shuxianbi@mail.ustc.edu.cn
{wenjiewang96, fulifeng93}@gmail.com
{mouyueqi, wangchenxu13, biao.tang, yanpeng04}@meituan.com

Abstract

Related search query recommendation is essential for enhancing user engagement and information discovery on digital platforms. While Large Language Models (LLMs) have shifted the field toward generative retrieval, existing methods suffer from two primary limitations: (1) pointwise generation via beam search often leads to semantic redundancy and wasted retrieval quota, and (2) current listwise approaches lack explicit reasoning, relying on superficial click-through rate (CTR) rewards. In this paper, we propose ReList, a novel framework that transforms related search into a reasoning-enhanced listwise generation task. ReList follows a two-stage training paradigm: first, Reasoning Activation constructs a high-quality dataset by back-translating diverse query lists into Chain-of-Thought (CoT) rationales; second, Alternative Training iteratively evolves the model using Reinforcement Learning with a Gated Multi-Objective Reward and a Corrective SFT mechanism to handle hard samples. Experimental results on real-world search benchmarks and online A/B tests demonstrate that ReList significantly outperforms state-of-the-art methods in both query diversity and user engagement, providing more insightful and logically grounded query recommendations.

1 Introduction

Related search query recommendations represent a fundamental component of modern digital platforms, serving to enhance user engagement and facilitate information discovery. As Figure 1 shows, the emergence of LLM has catalyzed a shift toward generative retrieval approaches for this task, which have demonstrated potential in generating contextually relevant and novel queries (Bi et al., 2025; Shao et al., 2025) to provoke users' search intents.

Despite their initial success, existing generative retrieval paradigms for related search suffer from two critical limitations. First, most current methods rely on a *pointwise generation* strategy via beam search (Bi et al., 2025; Deng et al., 2025; Rajput et al., 2023; Bao et al., 2025), which often leads to severe semantic redundancy. As illustrated in Figure 2 (A), these models frequently produce multiple candidates sharing identical prefixes or synonymous meanings, thereby squandering the limited retrieval quota and failing to cover a diverse range of user interests. Second, although recent efforts (Yin et al., 2025; Min et al., 2025; Xu et al., 2026; Liu et al., 2023) have attempted to generate the entire query list directly (*i.e.*, listwise generation), they still *lack explicit reasoning* capabilities during the generative process. These models typically employ a "black-box" mapping from user context to query sequences, primarily guided by surface-level rewards such as click-through rates (CTR). Consequently, they struggle to provide logically coherent and exploratory suggestions that can effectively guide users toward deeper information discovery.

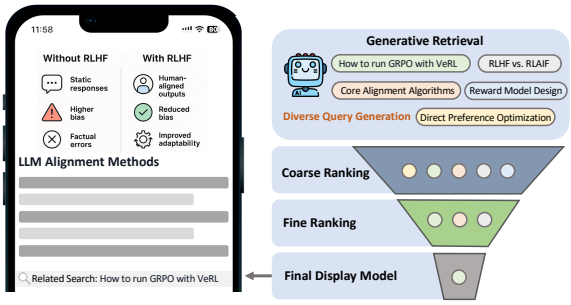

Figure 1: The architecture of our generative retrieval system for related search.

To address these issues, we propose **ReList**, a framework that transforms related search into a **R**easoning-**e**nhanced **List**wise generation task as Figure 2 (B) shows. Our approach consists of two key stages: (1) **Reasoning Activation**: We construct a specialized reasoning dataset by leverag-

*Work done during the internship at Meituan.
†Corresponding Authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1392–1405
July 2-7, 2026

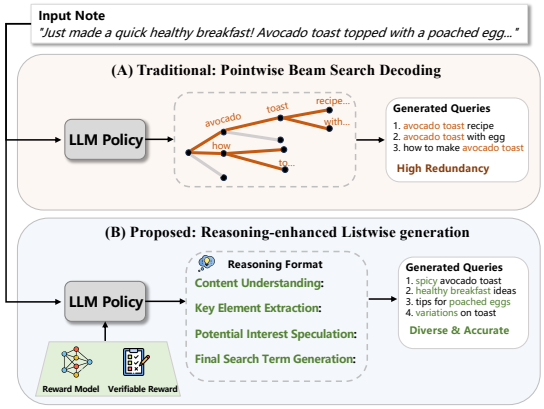

Figure 2: Motivation and illustration of our approach for diversified query generation. Traditional beam search decoding (A) produces multiple surface-level variations with high redundancy. In contrast, the proposed reasoning-based LLM policy trained with dual-reward reinforcement learning (B) captures underlying user intent and generates a set of diverse, accurate, and complementary search queries.

ing a teacher LLM to back-translate high-quality query lists—sampled via large beam search and diversity filtering—into Chain-of-Thought (CoT) rationales. This is further augmented with direct CoT-list synthesis to activate the model's reasoning ability. (2) **Alternative Training**: We iteratively evolve the model using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with a Gated Multi-Objective Reward mechanism to ensure stable optimization across diverse metrics. To handle hard samples, we introduce Corrective SFT, which utilizes a teacher LLM to rewrite suboptimal rollouts, subsequently filtering them through reward models to provide high-quality supervision. By integrating explicit reasoning with multi-objective reinforcement learning, ReList effectively moves beyond shallow pattern matching to provide logically coherent and diverse search suggestions.

The main contributions of this work are summarized as follows:

- We identify that existing generative retrieval suffers from severe semantic redundancy in pointwise approaches and a lack of explicit reasoning in current listwise paradigms.
- We propose ReList, a two-stage framework featuring Reasoning Activation via CoT synthesis and Alternative Training with Gated Multi-Objective rewards and Corrective SFT.
- Extensive experiments on real-world benchmarks and online A/B test demonstrate that ReList achieves SOTA performance in both query diversity and user engagement.

2 Related Work

Query generation on content platforms aims to synthesize search terms that resonate with a user's immediate interests (Li et al., 2024). Traditional techniques primarily focus on *active search* scenarios, where users have already initiated a query prefix. These methods encompass query suggestion (Wang et al., 2020; Bacciu et al., 2024; Guo et al., 2026), query rewriting (Feng et al., 2024; Peng et al., 2024), and personalized completion (Zhong et al., 2020; Wang et al., 2025; Li et al., 2020) based on historical interactions. While effective, these approaches assume an explicit search intent has already been demonstrated.

In contrast, our work focuses on the *discovery-oriented* setting, providing search options while users are browsing content to stimulate active exploration. While early studies utilized seq2seq architectures (Nogueira et al., 2019; Penha et al., 2023), recent advances have leveraged Large Language Models (LLMs) to generate queries from textual (Sannigrahi et al., 2024; Shao et al., 2025; Bi et al., 2025) or multimodal contexts (Wang et al., 2024). There are also other works adopt listwise generation paradigm to retrieval the whole query list at once (Yin et al., 2025; Min et al., 2025). However, existing generative retrieval paradigms often treat the process as a black-box mapping, overlooking the necessity of explicit reasoning and the challenges of multi-objective alignment. Most current methods struggle to balance CTR with exploratory diversity. **ReList** addresses this gap by introducing a reasoning-enhanced framework that optimizes for both engagement and intent expansion via a gated multi-objective reward mechanism.

3 Methodology

We present ReList, a two-stage framework designed for reasoning-enhanced listwise generation. The process begins with **Reasoning Activation** to establish reasoning ability, followed by an **Alternative Training** stage that employs multi-objective RL and corrective SFT to refine the output.

3.1 Problem Formulation

The objective is to learn a conditional probability distribution $P_\theta(y, z|x)$, where x denotes the input user context and the target output consists of a la-

tent reasoning state z (e.g., Chain-of-Thought) and a sequence of diverse queries $y = [q_1, q_2, \ldots, q_k]$. Unlike conventional pointwise methods, ReList explicitly models the generation process by decomposing the joint distribution:

$$P_\theta(y, z|x) = P_\theta(z|x) \prod_{i=1}^{k} P_\theta(q_i|x, z, q_{<i}), \quad (1)$$

where $q_{<i}$ represents the queries previously generated within the list. This formulation ensures that each query q_i is conditioned not only on the input x and the reasoning rationale z, but also on the preceding candidates to minimize semantic redundancy. The framework is optimized via a two-stage curriculum:

Reasoning Activation Initializing the latent distribution $P_\theta(y, z|x)$ to align the model's internal logic with diverse search intents.

Alternative Training Iteratively refining the policy to maximize a multi-objective reward function $\mathcal{R}(x, y, z)$ that balances multiple objectives such as relevance, diversity, and reasoning quality.

3.2 Stage I: Reasoning Activation

In this stage, we construct a high-quality reasoning-to-list dataset through a combination of candidate distillation and direct knowledge synthesis.

3.2.1 Candidate Sampling

We begin by training a base pointwise model θ_{pt} using standard SFT and Direct Preference Optimization (DPO) (Rafailov et al., 2023) on historical click-through data. To generate diverse candidates, we employ Beam Search on θ_{pt} to produce an initial pool of potential queries $\mathcal{Q} = \{q'_1, q'_2, \ldots, q'_m\}$ with a large m. To transform these independent candidates into an optimal list y, we utilize a teacher LLM (LLM_t) to act as a diversity-aware reranker. Specifically, given an instruction prompt $\mathcal{P}_s$ for selection, the target list is generated via $y^* \sim P_{\text{LLM}_t}(y \mid \mathcal{P}_s, x, \mathcal{Q})$. The teacher model selects and orders k queries from $\mathcal{Q}$ that maximize both semantic coverage and intent relevance, resulting in the final target list y^*.

3.2.2 CoT Induction

Since raw search logs lack explicit reasoning, we employ LLM_t to induce the latent state z via back-translation. Given the input context x and the refined list y^*, the teacher model is prompted with a specific instruction template $\mathcal{P}_l$ to generate a Chain-of-Thought rationale as $z^* \sim P_{\text{LLM}_t}(z \mid \mathcal{P}_b, x, y^*)$, where z^* explains the underlying search motivations and the logical transitions between queries in y^*. This process ensures that every target list is grounded in a coherent reasoning path.

3.2.3 Dataset Composition

To prevent the model from inheriting repetitive patterns or biases from the pointwise generator, we further augment the training set with *Direct Synthesis*. We prompt LLM_t with an augmentation prompt $\mathcal{P}_a$ to directly generate pairs $(z, y) \sim P_{\text{LLM}_t}(z, y \mid \mathcal{P}_a, x)$ from scratch for a subset of contexts x. This provides the model with "silver-standard" examples that exhibit higher linguistic variety and more complex reasoning structures than those derived from historical logs. The final activation dataset $\mathcal{D}_{\text{act}}$ is formed by the union of these two parts: $\mathcal{D}_{\text{act}} = \{(x, z^*, y^*)_{\text{sampled}}\} \cup \{(x, z, y)_{\text{synth}}\}$ We then perform supervised fine-tuning on our policy θ using $\mathcal{D}_{\text{act}}$ to activate its ability to "think" before suggesting.

3.3 Stage II: Alternative Training

To further refine the model's ability to generate diverse and logically grounded query lists, we propose an Alternative Training paradigm. This stage proceeds for T iterations, where each iteration $t \in \{1, \ldots, T\}$ consists of two alternating phases: Multi-objective Reinforcement Learning to explore the policy space, and Corrective SFT to stabilize the model on hard samples. Formally, the transition from policy θ_t to θ_{t+1} is defined as:

$$\theta_{t+1} = \text{SFT}\left(\text{RL}(\theta_t, \mathcal{R}), \mathcal{D}^{(t)}_{\text{corr}}\right), \quad (2)$$

where $\mathcal{R}$ denotes the composite reward and $\mathcal{D}^{(t)}_{\text{corr}}$ is the corrective dataset synthesized in round t.

3.3.1 Multi-objective Reinforcement Learning

In the RL phase, we utilize Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to optimize the policy. For a given context x, we sample a group of G outputs $\{(z_i, y_i)\}_{i=1}^{G}$ from the current policy π_{θ_t}.

Gated Multi-Objective Reward To ensure structural compliance and high performance, we design a Gated Multi-Objective Reward mechanism. A gating function $\Phi_{\text{gate}}(x, y, z) \in \{0, 1\}$ acts as a hard constraint filter: $\Phi_{\text{gate}} = \mathbb{I}_{\text{xml}} \cdot \mathbb{I}_{\text{fmt}} \cdot \mathbb{I}_{\text{rel}}$, where $\mathbb{I}_{\text{xml}}$ validates the presence of `<think>`

and <answer> tags, $\mathbb{I}_{\text{fmt}}$ checks query counts and length constraints, and $\mathbb{I}_{\text{rel}}$ is a binary signal from a pre-trained relevance model requiring $\min_{q \in y} \text{Rel}(x, q) > \tau_{\text{rel}}$. If any hard constraint is violated, the total reward is nullified.

The soft rewards $\mathcal{R}_{\text{soft}}$ consist of:

- **Thinking Length Reward** (r_{len}): Encourages sufficient reasoning. $r_{\text{len}} = 1$ if $\text{len}(z) \in [L_1, L_2]$, and 0 otherwise.
- **Reference Reward** (r_{ref}): Measures alignment with high-CTR historical queries. Given a reference query q_{ref}, $r_{\text{ref}} = \max_{q_i \in y} \cos(\mathbf{e}_{q_i}, \mathbf{e}_{q_{\text{ref}}})$, where $\mathbf{e}$ is the text embedding.
- **CTR Reward** (r_{ctr}): Estimated via a Bradley-Terry reward model trained on historical click data. We compute $r_{\text{ctr}} = \text{READOUT}(\{\sigma(r_1^{\text{ctr}}), \ldots, \sigma(r_k^{\text{ctr}})\})$ to represent the collective engagement potential of the list. Since the raw output of the Bradley-Terry reward model lies in $\mathbb{R}$, directly using it leads to severe training instability; we therefore apply the sigmoid function $\sigma(\cdot)$ to normalize each per-query score so that its scale is consistent with the other reward terms.
- **Diversity Reward** (r_{div}): Penalizes semantic overlap within the list:

$$r_{\text{div}} = 1 - \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \cos(\mathbf{e}_{q_i}, \mathbf{e}_{q_j}). \tag{3}$$

The final composite reward is $\mathcal{R} = \Phi_{\text{gate}} \cdot \sum_k w_k r_k$. Under this signal, the GRPO loss for round t is:

$$\mathcal{L}_{\text{RL}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \min\big(\rho_i \hat{A}_i,\ \rho_i^{\text{clip}} \hat{A}_i\big) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \Bigg], \tag{4}$$

where $\rho_i = \frac{\pi_\theta(y_i, z_i|x)}{\pi_{\theta_t}(y_i, z_i|x)}$, $\rho_i^{\text{clip}} = \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)$, and $\hat{A}_i = \frac{\mathcal{R}_i - \text{mean}(\{\mathcal{R}\}_{j=1}^G)}{\text{std}(\{\mathcal{R}\}_{j=1}^G)}$ represents the advantage computed by group-wise reward normalization.

3.3.2 Corrective SFT

While RL excels at exploring the average case, it often fails to optimize "hard" long-tail samples where the initial policy π_θ cannot find a high-reward trajectory. To address this, we introduce Corrective SFT. For any context x where all sampled rollouts in the RL phase fail to meet a quality threshold (i.e., $\max(\{R_i\}_{i=1}^G) < \tau_l$), we utilize the teacher model LLM_t to perform a corrective rewrite. Given a correction prompt $\mathcal{P}_c$, the teacher generates a new reasoning-list pair: $(z', y') \sim P_{\text{LLM}_t}(z, y \mid \mathcal{P}_c, x, \{z_i, y_i\}_{i=1}^G)$. The teacher model analyzes the failed rollouts to identify errors and produces a corrected version. If the rewritten pair satisfies $R(x, z', y') > \tau_h$, it is added to a corrective dataset $\mathcal{D}_{\text{corr}}^{(t)}$. At the end of each round t, the policy θ is updated via SFT on $\mathcal{D}_{\text{corr}}^{(t)}$, allowing the model to learn from the "rectified" logic of these challenging cases:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,z',y') \in \mathcal{D}_{\text{corr}}^{(t)}} [\log P_\theta(z', y' \mid x)]. \tag{5}$$

4 Experiments

4.1 Experimental Setup

Dataset While several public datasets contain related search scenarios (e.g., GREAT (Shao et al., 2025), QiLin (Chen et al., 2025)), they typically lack CTR metrics, which are essential for our reward-driven training objectives. Consequently, we collected a large-scale dataset from a leading content platform, consisting of approximately 1.5 million (note, query) pairs for offline experiments. For evaluation, we sampled 2,000 notes as a test set using a category-balanced sampling strategy to ensure a comprehensive assessment across topics.

Baselines We compare ReList against two groups of baselines. First, we evaluate state-of-the-art closed-source LLMs in a zero-shot listwise generation setting, including GLM-4.5 (GLM, 2025), GPT-4o, Doubao-1.5-pro (Seed, 2025), Qwen-Max (Yang et al., 2025a), and Deepseek-Chat (DeepSeek-AI, 2025). Second, we utilize Qwen2.5-7B-Instruct (Yang et al., 2025b) as the open-source backbone to evaluate various retrieval paradigms: (i) *Zero-Shot* (Beam Search, Diverse Beam Search, and Listwise Generation), (ii) *SFT* (Beam Search and Diverse Beam Search), and (iii) *DPO* (Beam Search and Diverse Beam Search). The DPO-tuned model also serves as the pointwise generator θ_{pt} for our reasoning activation stage.

Evaluation Metrics Following the generative retrieval setting, each method is required to generate

Table 1: Comparison of different methods. **Bold** indicates the best result, and underline indicates the second best.

Method	Variant	CTR Score			Div.	Rel.	Reference Match		
		Mean ↑	Max ↑	Min ↑	Mean ↑	Mean ↑	Edit Dist. ↓	BLEU ↑	Sim. ↑
Backbone: Closed-source LLMs									
Zero-Shot	GLM-4.5	-2.116	0.484	-4.214	0.500	0.839	0.713	0.290	0.660
	GPT-4o	-2.430	-0.437	-4.129	**0.515**	0.783	0.825	0.185	0.584
	Doubao-1.5-pro	**-1.297**	**2.738**	-4.011	0.436	**0.908**	**0.667**	**0.339**	**0.697**
	Qwen-Max	-2.083	0.580	-4.153	0.506	0.849	0.717	0.298	0.664
	Deepseek-Chat	-1.663	1.288	**-3.878**	0.489	0.853	0.704	0.311	0.673
Backbone: Qwen2.5-7B-Instruct									
Zero-Shot	Beam Search	-2.093	0.089	-4.158	0.302	0.911	0.649	0.296	0.690
	Diverse Beam Search	-1.981	-0.407	-3.534	0.224	**0.916**	0.680	0.263	0.668
	Listwise Generation	-2.697	-0.314	-4.734	0.522	0.814	0.760	0.228	0.615
SFT	Beam Search	1.607	3.446	-0.391	0.247	0.890	**0.436**	**0.549**	**0.811**
	Diverse Beam Search	1.463	3.330	-0.613	0.279	0.888	0.442	0.539	0.810
DPO	Beam Search	3.674	5.131	2.019	0.167	0.889	0.504	0.494	0.775
	Diverse Beam Search	3.635	5.347	2.358	0.219	0.868	0.509	0.486	0.774
ReList	Listwise Generation	**5.107**	**6.886**	**2.845**	**0.566**	0.883	0.603	0.392	0.720

a list of 5 queries for each input note. We evaluate performance across three dimensions:

- **CTR Score**: We report the Mean, Max, and Min CTR of the generated query list to reflect the potential user engagement. In the offline evaluation stage, we utilize the CTR reward model to reflect the CTR Score.
- **Diversity (Div.)**: Calculated as the mean pairwise semantic distance (1 − cosine similarity) within the generated list, which is similar to the diversity reward.
- **Relevance (Rel.)**: Measured by the mean semantic similarity between the generated queries and the input context, implemented by the trained relevance reward model.
- **Reference Match**: To assess alignment with real-world user intent, we compare outputs against high-CTR historical queries using Edit Distance, BLEU, and Semantic Similarity (Sim.) evaluated by text embedding.

Implementation Details Our framework uses Qwen2.5-7B-Instruct as the backbone policy model. The SFT baselines are fine-tuned for 3 epochs on the full dataset. For the DPO baselines, we perform iterative DPO to optimize the CTR reward. During the Reasoning Activation stage, we synthesize approximately 180,000 samples, maintaining a 2:8 ratio between *Direct Synthesis* and *Pointwise-based Refinement* to align with the online query distribution. For Alternative Training, we use 20,000 notes and set $T = 5$ iterations. The CTR reward readout function is set to *mean*, and all reward weights are set to $w_i = 1$. We use BGE-large-zh-1.5 (Xiao et al., 2023) for all text embeddings and GPT-4o as the teacher model LLM_t. Training is implemented using the `trl` framework for SFT, DPO and the `verl` framework for RL.

4.2 Main Results

Results in Table 1 yield the following key insights:

Diversity vs. Engagement Listwise generation inherently achieves superior diversity compared to pointwise methods. This is because listwise modeling accounts for inter-query dependencies, effectively minimizing semantic overlap. However, without domain-specific optimization, zero-shot models suffer from poor user engagement, as evidenced by their negative CTR scores.

Pointwise Limitations While pointwise SFT and DPO methods significantly improve CTR scores, they struggle to optimize diversity directly. Because candidates are generated independently, these models often produce homogenized results (Div. < 0.3), failing to adequately cover the diverse spectrum of potential user search intents.

ReList Superiority Our framework, ReList, achieves the best of both worlds. By integrating reasoning-enhanced listwise generation with

Table 2: Ablation study of **ReList** on the test set.

Config.	CTR ↑	Div. ↑	Rel. ↑
Full Model	5.1070	0.5669	0.8836
w/o CTR	-1.1050	0.6545	0.8921
w/o Div	**6.2590**	0.1505	**0.9273**
w/o Rel	4.5049	**0.7041**	0.5291
w/o Stage I	-0.0632	0.5823	0.8591
w/o Stage II	1.7683	0.3313	0.9081
w/o C-SFT	4.7534	0.3711	0.9150
w/o CoT	4.8450	0.5455	0.8856

multi-objective reinforcement learning, it delivers SOTA CTR (5.1070) while reaching peak diversity (0.5669). These results demonstrate that ReList not only captures user intent more accurately than traditional alignment methods but also provides a more exploratory and varied search experience.

4.3 Ablation Study

We conduct an ablation study to evaluate the contribution of each reward component and training stage, as summarized in Table 2.

Impact of Multi-objective Rewards Removing the CTR or Diversity reward leads to a catastrophic drop in its respective metric. This confirms that the Gated Multi-Objective mechanism is essential for balancing these competing goals, preventing the model from sacrificing intent coverage for engagement or vice versa.

Necessity of Training Stages Stage II (Alternative Training) is the most critical component; its removal causes the CTR to decrease, proving that iterative RL is the primary driver of performance. Additionally, omitting Stage I (Reasoning Activation) or Corrective SFT leads to a significant decrease in diversity, highlighting their roles in establishing a diverse policy starting point and handling long-tail samples.

Role of Reasoning The removal of CoT results in a decline in both CTR and Diversity. This suggests that explicit reasoning rationales serve as a necessary logical anchor, helping the model maintain coherence and explore distinct search intents more effectively.

4.4 Impact of Iteration Rounds

To validate the necessity of the multi-round Alternative Training and provide empirical guidance for hyperparameter selection, we vary the number of iteration rounds T from 0 to 5 on the Qwen2.5-7B-Instruct backbone while fixing all other hyperparameters. The trends of CTR score and Diversity are shown in Figure 3.

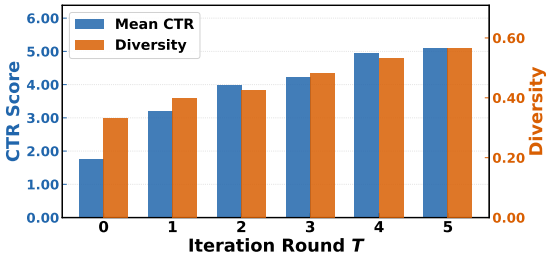

Figure 3: Impact of iteration rounds T on CTR score and Diversity.

Sustained Improvement on Both Objectives Both CTR and Diversity exhibit a monotonically increasing trend with T. At $T = 0$ (i.e., after only Stage I without any RL iteration), the model obtains a CTR of 1.77 and Diversity of 0.33, confirming that while reasoning activation lays the foundation, alignment with user engagement and diverse intent coverage still requires iterative refinement in Stage II.

Diminishing Marginal Gains Early iterations yield the largest gains: CTR rises from 1.77 to 3.20 as T goes from 0 to 1, and further to 3.98 at $T = 2$. In contrast, CTR only slightly increases from 4.96 to 5.11 between $T = 4$ and $T = 5$, and Diversity grows from 0.53 to 0.57, indicating that the policy asymptotically approaches the attainable frontier under the current reward.

Synchronous Pareto Progress Throughout the entire iteration process, CTR and Diversity improve *simultaneously*, without the "seesaw" effect observed in the ablation (e.g., removing the Diversity reward boosts CTR but collapses Diversity). This confirms that the Gated Multi-Objective mechanism consistently drives training along a Pareto-optimal frontier. Given the saturation at $T = 5$, we adopt $T = 5$ as the default configuration.

4.5 Reward Configuration Analysis

We analyze the sensitivity of the CTR reward weight w_{ctr} and the design choices of the readout function and value-range normalization, as summarized in Table 3.

Table 3: Sensitivity of reward weight w_{ctr} and effects of readout / normalization choices.

Config.	CTR ↑	Div. ↑	Rel. ↑
$w_{\text{ctr}} = 0.1$	-0.103	0.628	0.900
$w_{\text{ctr}} = 0.5$	4.156	0.528	0.894
$w_{\text{ctr}} = 1.0$	5.107	0.567	0.884
$w_{\text{ctr}} = 1.5$	5.401	0.492	0.887
$w_{\text{ctr}} = 2.0$	5.697	0.480	0.881
$w_{\text{ctr}} = 3.0$	6.049	0.454	0.878
Readout = max	0.416	0.623	0.904
w/o Normalization	9.211	0.017	0.931

CTR-Diversity Trade-off As w_{ctr} increases from 0.1 to 3.0, CTR rises monotonically, while Diversity decreases. An overly small weight (0.1) weakens the CTR signal and degrades the model into a diversity-driven generator that loses fit to user preferences, whereas an overly large weight suppresses intent coverage by over-pursuing high-click phrases. $w_{\text{ctr}} = 1.0$ achieves the best Pareto balance (CTR 5.107, Div. 0.567) and is adopted as the default.

Readout Function Replacing the mean-based aggregation of r_{ctr} with the max operator sharply reduces CTR to 0.416. Max aggregation only rewards the single best query in the list, while leaving the remaining queries unconstrained, which encourages an unbalanced output of "one strong query plus several weak ones". Mean aggregation instead imposes a uniform quality constraint across the entire list, aligning naturally with the diversity objective.

Value-Range Normalization Disabling the Sigmoid-based normalization on r_{ctr} leads to a catastrophic mode collapse: CTR surges to 9.211 while Diversity vanishes to 0.017. Since the raw CTR reward is unbounded in $\mathbb{R}$ while the diversity reward is naturally bounded in $[0, 1]$, without normalization the CTR gradient dominates and the model degenerates into a pure click-rate maximizer. Applying Sigmoid to compress r_{ctr} into $(0, 1)$ aligns the numerical magnitudes of all rewards and enables the Gated Multi-Objective mechanism to effectively coordinate competing objectives.

4.6 Online A/B Test

To verify the effectiveness of ReList in production, we conducted a two-week online A/B test on a prominent local lifestyle platform involving 15 million content items.

Table 4: The relative improvement of online A/B test.

Online Metrics	Ours
Passive Search QV	+0.930%
Search QV Ratio (Guidance)	+1.050%
PV CTR (Search Guidance)	+1.045%
Passive Search UV	+0.745%
Passive Search per User	+0.192%
Query Change Rate	-0.232%

Experimental Setup We integrated ReList as a supplementary retrieval pathway using an offline/nearline pre-computation strategy. By caching query lists for new content in advance, we satisfied strict industrial latency requirements while eliminating real-time LLM inference overhead.

Results Analysis As shown in Table 4, ReList yielded significant improvements across key business metrics:

- **Engagement**: PV CTR and Passive Search QV increased by **+1.045%** and **+0.930%** respectively, reflecting the model's superior ability to capture user intent.
- **User Reach**: Passive Search UV grew by **+0.745%**, indicating a broader positive impact on the user base.
- **Efficiency**: The Search QV Ratio improved by **+1.050%** while the Query Change Rate dropped by **-0.232%**, suggesting that users find relevant information more directly with fewer reformulations.

5 Conclusion

We propose ReList, a reasoning-enhanced listwise generation framework that overcomes semantic redundancy and reasoning gaps in related search. By integrating reasoning activation with iterative multi-objective reinforcement learning, our approach transforms query suggestion into a logically grounded discovery process. Extensive offline experiments and large-scale online A/B tests demonstrate that ReList significantly outperforms state-of-the-art baselines in both engagement and diversity, confirming its practical effectiveness in real-world industrial settings.

Acknowledgments

This research was supported by Meituan, National Natural Science Foundation of China (U25B2071), and the advanced computing resources provided by the Supercomputing Center of the USTC.

Ethical Considerations

Our generative query recall mechanism adheres to two primary ethical pillars: **(1) Privacy & Anonymization**: We employ a zero-linkage policy, utilizing strictly anonymized, aggregated search sessions. No user-specific profiles, search histories are accessed, ensuring generated queries remain entirely dissociated from individual identities. **(2) Bias & Safety Mitigation**: To prevent the propagation of latent biases, we implement a multi-layer defense: automated content filtering, proactive exclusion of sensitive socio-political topics during generation, and regular human audits to identify and rectify non-obvious harmful associations.

References

Andrea Bacciu, Enrico Palumbo, Andreas Damianou, Nicola Tonellotto, and Fabrizio Silvestri. 2024. Generating query recommendations via llms. *CoRR*, abs/2405.19749.

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. *Trans. Recomm. Syst.*, 3(4):53:1–53:27.

Shuxian Bi, Chongming Gao, Wenjie Wang, Yueqi Mou, Chenxu Wang, Tang Biao, Peng Yan, and Fuli Feng. 2025. Consistency-aware online multi-objective alignment for related search query generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1365–1377.

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, Qingyao Ai, and Yiqun Liu. 2025. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 3670–3680. ACM.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *CoRR*, abs/2502.18965.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9571–9583. Association for Computational Linguistics.

GLM. 2025. GLM-4.5: agentic, reasoning, and coding (ARC) foundation models. *CoRR*, abs/2508.06471.

Xian Guo, Ben Chen, Siyuan Wang, Ying Yang, Mingyue Cheng, Chenyi Lei, Yuqing Ding, and Han Li. 2026. Onesug: The unified end-to-end generative framework for e-commerce query suggestion. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 14774–14782. AAAI Press.

Ying Li, Jizhou Huang, Miao Fan, Jinyi Lei, Haifeng Wang, and Enhong Chen. 2020. Personalized query auto-completion for large-scale POI search at baidu maps. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(5):70:1–70:16.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024. A survey of generative search and recommendation in the era of large language models. *CoRR*, abs/2404.16924.

Shuchang Liu, Qingpeng Cai, Zhankui He, Bowen Sun, Julian J. McAuley, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Generative flow network for listwise recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1524–1534. ACM.

Erxue Min, Hsiu-Yuan Huang, Xihong Yang, Min Yang, Xin Jia, Yunfang Wu, Hengyi Cai, Junfeng Wang, Shuaiqiang Wang, and Dawei Yin. 2025. Ctr-guided generative query suggestion in conversational search. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 2624–2634. Association for Computational Linguistics.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *CoRR*, abs/1904.08375.

Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024. Large language model based long-tail query rewriting in taobao search. In *Companion*

Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, pages 20–28. ACM.

Gustavo Penha, Enrico Palumbo, Maryam Aziz, Alice Wang, and Hugues Bouchard. 2023. Improving content retrievability in search with controllable query generation. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3182–3192. ACM.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Sonal Sannigrahi, Thiago Fraga-Silva, Youssef Oualil, and Christophe Van Gysel. 2024. Synthetic query generation using large language models for virtual assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2837–2841. ACM.

ByteDance Seed. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *CoRR*, abs/2504.13914.

Ninglu Shao, Jinshan Wang, Chenxu Wang, Qingbiao Li, and Xiaoxue Zang. 2025. GREAT: guiding query generation with a trie for recommending related search about video at kuaishou. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 4818–4826. ACM.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Sida Wang, Weiwei Guo, Huiji Gao, and Bo Long. 2020. Efficient neural query auto completion. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2797–2804. ACM.

Zheng Wang, Bingzheng Gan, and Wei Shi. 2024. Multimodal query suggestion with multi-agent reinforcement learning from human feedback. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1374–1385. ACM.

Zhibo Wang, Xiaoze Jiang, Zhiheng Qin, and Enyun Yu. 2025. Personalized query auto-completion for long and short-term interests with adaptive detoxification generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 5018–5028. ACM.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Jingcao Xu, Jianyun Zou, Renkai Yang, Zili Geng, Qiang Liu, and Haihong Tang. 2026. AIGQ: an end-to-end hybrid generative architecture for e-commerce query recommendation. *CoRR*, abs/2603.19710.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025b. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Junhao Yin, Haolin Wang, Peng Bao, Ju Xu, and Yongliang Wang. 2025. From clicks to preference: A multi-stage alignment framework for generative query suggestion in conversational system. *CoRR*, abs/2508.15811.

Jianling Zhong, Weiwei Guo, Huiji Gao, and Bo Long. 2020. Personalized query suggestions. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1645–1648. ACM.

Appendix

A More Details on the Reward Models

CTR Reward Modeling To accurately differentiate user engagement levels, we developed a CTR Reward Model designed to estimate the relative click-through potential between query candidates for a shared context. This model incorporates a linear regression layer atop the Qwen2.5-1.5B backbone architecture. We constructed the training corpus by filtering content-query interactions with a minimum of 100 impressions. To mitigate noise, we applied z-tests to the underlying click distributions, preserving only pairs with statistically significant CTR differences ($p < 0.01$). This rigorous selection process resulted in a dataset of approximately 330,000 preference triplets $(\mathbf{x}, \mathbf{y}_+, \mathbf{y}_-)$, where $\mathbf{y}_+$ exhibits a demonstrably higher engagement rate than $\mathbf{y}_-$ for the context $\mathbf{x}$. The model is optimized using the Bradley-Terry objective (6), ensuring the learned reward scalar r_{ctr} effectively captures the probability of user preference:

$$\mathcal{L}_{\text{BT}} = -\log \sigma \left(r_{\text{ctr}}(\mathbf{x}, \mathbf{y}+) - r_{\text{ctr}}(\mathbf{x}, \mathbf{y}_-) \right). \tag{6}$$

Relevance Reward Model To enforce the hard relevance constraint $\mathbb{I}_{\text{rel}}$ within the gating mechanism, we developed a binary classification model to evaluate the semantic alignment between the input context x and each generated query q_i. The model was trained on a dataset of 340,000 context-query pairs. Given the significant class imbalance between positive and negative samples in the raw search logs, we employed a downsampling strategy to ensure a balanced training distribution. We utilized the `Qwen2.5-1.5B` architecture as the backbone, fine-tuned specifically as a sequence classifier. The final relevance model achieved an F1 score of 0.9497, providing a highly reliable signal for the hard-gating mechanism during the reinforcement learning phase.

B Complexity Analysis

To evaluate the practical feasibility of ReList, we analyze its inference complexity in comparison with traditional Beam Search based pointwise retrieval. We assume both methods share the same prefill complexity and focus on the decoding phase.Let k be the number of target queries, l be the average length of each individual query, and m be the length of the reasoning rationale z.

ReList Our method generates the reasoning path and the entire query list in a single autoregressive pass. The total number of decoding steps is $m + k \cdot l$, resulting in a complexity of $O(m + k \cdot l)$.

Beam Search : To obtain k high-quality and diverse candidates, pointwise methods typically require a beam size b (where $b \geq k$). To generate a set of candidates of length l, the decoding complexity is $O(b \cdot l)$.

Efficiency Gains ReList exhibits superior inference efficiency when the following condition is met: $m < (b - k) \cdot l$. In practical industrial applications, a large beam size (e.g., $b \gg k$) is often necessary to ensure sufficient candidate diversity and to mitigate the semantic redundancy inherent in pointwise generation. When b is large, the computational overhead of maintaining multiple hypotheses in beam search exceeds the cost of generating a single reasoning rationale m. Furthermore, since ReList generates the list as a continuous sequence, it avoids the redundant prefix computations and repetitive hidden-state evaluations that occur when beam search tracks multiple similar paths.

C Theoretical Analysis on Diversity

To provide a theoretical foundation for the proposed approach, we analyze the redundancy mechanism in traditional beam search relative to our listwise paradigm through the lens of distributional divergence.

Lipschitz Continuity of Generation. Let f_θ denote the LLM that maps a hidden state sequence h to a probability distribution over the vocabulary $\mathcal{V}$. It is assumed that f_θ satisfies the Lipschitz continuity condition with constant L:

$$D_{KL}(P(\cdot|h_i)||P(\cdot|h_j)) \leq L\|h_i - h_j\|, \tag{7}$$

where D_{KL} denotes the Kullback-Leibler divergence. This condition implies that hidden states in close geometric proximity yield highly similar distributions for the subsequent token.

Collapse in Beam Search. In beam search, k hypotheses $\{h_1, \ldots, h_k\}$ are maintained to maximize the cumulative log-probability. Due to the heavy-tailed nature of language modeling objectives, the top-k candidates frequently exhibit significant lexical overlap, such as shared prefixes or synonym substitutions. Consequently, the pairwise distance between the hidden states of these hypotheses is bounded by a small value ϵ: $\|h_i - h_j\| < \epsilon$.By applying the Lipschitz continuity condition, the divergence between the next-token distributions of these hypotheses approaches zero:

$$\lim_{\epsilon \to 0} D_{KL}(P(\cdot|h_i)||P(\cdot|h_j)) = 0. \tag{8}$$

This phenomenon, defined here as *Local Mode Collapse*, causes all beams to greedily sample from the same dominant mode, which inevitably results in semantically homogeneous queries.

Diversity via Conditional Mode Shifting. In contrast, the proposed listwise framework generates queries sequentially as a single chain $Y = [q_1, q_2]$. The generation of the t-th token of q_2 is conditioned on the complete trajectory of q_1. This process introduces a significant semantic shift vector Δ to the hidden state: $h_{new} = h_{old} + \Delta(q_1)$.When the model is trained to penalize redundancy through reinforcement learning rewards, this shift functions as a repulsive potential that forces the distribution $P(\cdot|x, q_1)$ to suppress the modes already covered by q_1. This mechanism, termed *Conditional Mode Shifting*, theoretically ensures that subsequent queries are sampled from unexplored regions of the intent spectrum.

D Performance on Different Backbones

To investigate the scalability and robustness of our framework, we evaluate ReList across three model scales: Qwen2.5-1.5B, 3B, and 7B. The results, summarized in Table 5, reveal several key findings:

Positive Scaling Trend We observe a consistent performance gain as the model parameter count increases. Specifically, the Mean CTR of ReList rises from 4.548 (1.5B) to 4.823 (3B), and finally to 5.107 (7B). A similar upward trend is observed in diversity metrics, with BGE scores improving from 0.516 to 0.566. This suggests that larger backbones possess stronger latent reasoning capabilities, which our framework effectively leverages to generate higher-quality search intents.

Consistent Superiority across Scales Regardless of the model size, ReList consistently outperforms all baseline paradigms, including SFT and DPO-tuned pointwise models. Notably, even our smallest variant (ReList-1.5B) achieves a Mean CTR of 4.548, which significantly surpasses the performance of the largest pointwise baseline (DPO-7B, CTR: 3.674). This highlights the efficiency of our reasoning-enhanced listwise paradigm, demonstrating that logical grounding is a more potent driver of performance than raw parameter scale alone.

Robustness of Listwise Modeling While smaller zero-shot models often struggle with engagement (e.g., negative CTR scores for Qwen2.5-1.5B/3B), the application of our two-stage training paradigm successfully activates their potential. Even at smaller scales, the model maintains a high diversity score (above 0.51), confirming that the listwise generation approach remains a robust solution for mitigating redundancy, regardless of the underlying backbone's capacity.

E Prompts

We list the four prompts used in the ReList pipeline: $\mathcal{P}_s$ (reranker, constructing target lists from pointwise candidates), $\mathcal{P}_l$ (CoT reasoning induction, synthesizing rationales), $\mathcal{P}_a$ (listwise query generation, used both for direct synthesis and at inference), and $\mathcal{P}_c$ (correction, used in Corrective SFT).

Table 5: Comparison of different methods. **Bold** indicates the best result, and underline indicates the second best.

Method	Variant	CTR Score			Div.	Rel.	Reference Match		
		Mean	Max	Min	BGE	Score	Edit Dist. ↓	BLEU ↑	Sim. ↑
Backbone: Qwen2.5-1.5B-Instruct									
Zero-shot	Beam Search	-0.902	0.733	-2.497	0.344	0.713	0.773	0.263	0.647
	Diverse Beam Search	-0.335	1.549	-2.153	0.292	0.793	0.779	0.249	0.657
	Listwise Generation	-1.197	1.141	-3.271	0.363	0.732	0.800	0.197	0.578
SFT	Beam Search	1.843	3.569	-0.051	0.223	**0.906**	**0.490**	**0.499**	0.782
	Diverse Beam Search	1.692	3.457	-0.279	0.266	0.898	0.491	0.493	**0.785**
DPO	Beam Search	2.637	4.214	0.979	0.219	0.890	0.507	0.488	0.771
	Diverse Beam Search	2.522	4.142	0.740	0.270	0.883	0.503	0.490	0.779
ReList	Listwise Generation	**4.548**	**6.450**	**2.360**	**0.516**	0.858	0.632	0.372	0.715
Backbone: Qwen2.5-3B-Instruct									
Zero-shot	Beam Search	1.207	2.381	-0.037	0.132	**0.934**	0.774	0.228	0.668
	Diverse Beam Search	1.128	2.445	-0.275	0.164	0.934	0.777	0.220	0.673
	Listwise Generation	-2.018	0.432	-4.232	0.483	0.798	0.758	0.232	0.619
SFT	Beam Search	1.859	3.601	0.000	0.217	0.907	0.488	**0.498**	0.781
	Diverse Beam Search	1.746	3.446	-0.114	0.247	0.903	**0.486**	0.494	**0.783**
DPO	Beam Search	3.083	4.656	1.469	0.202	0.891	0.514	0.479	0.764
	Diverse Beam Search	2.968	4.513	1.357	0.242	0.885	0.507	0.483	0.772
ReList	Listwise Generation	**4.823**	**6.668**	**2.687**	**0.533**	0.842	0.639	0.363	0.717
Backbone: Qwen2.5-7B-Instruct									
Zero-shot	Beam Search	-2.093	0.089	-4.158	0.302	0.911	0.649	0.296	0.690
	Diverse Beam Search	-1.981	-0.407	-3.534	0.224	**0.916**	0.680	0.263	0.668
	Listwise Generation	-2.697	-0.314	-4.734	0.522	0.814	0.760	0.228	0.615
SFT	Beam Search	1.607	3.446	-0.391	0.247	0.890	**0.436**	**0.549**	**0.811**
	Diverse Beam Search	1.463	3.330	-0.613	0.279	0.888	0.442	0.539	0.810
DPO	Beam Search	3.674	5.131	2.019	0.167	0.889	0.504	0.494	0.775
	Diverse Beam Search	3.635	5.347	2.358	0.219	0.868	0.509	0.486	0.774
ReList	Listwise Generation	**5.107**	**6.886**	**2.845**	**0.566**	0.883	0.603	0.392	0.720

Reranker Prompt $\mathcal{P}_s$

Role: Query optimization expert that selects and complements queries from a candidate pool with balanced diversity, relevance, and appeal.
Input: Note information ({{title, content, shopinfo}}); Candidate pool of 20 queries generated by the pointwise model ({{candidate_list}}).
Task: (i) *Select* the best queries using criteria of diversity (non-overlapping topics), relevance (no hallucination/typos), and appeal (click-inducing); (ii) *Complete* by regenerating new queries when candidate quality is insufficient, ensuring a final total of **5 queries**.
Output (strict JSON): `{"final_keywords":[q1..q5], "newly_generated_keywords":[...]}`; set the second field to `[]` if no new queries are added.

CoT Reasoning Induction Prompt $\mathcal{P}_l$

Role: Search reasoning expert that explains how note content inspires a related query set with a logically rigorous thought process.
Input: {{title}}, {{content}}, {{shopinfo}}, {{final_keywords}}.
Task: Produce a natural, coherent Chain-of-Thought ($\leq$500 words) covering four stages: (1) *Content Understanding* – theme, key terms, emotional tone; (2) *Interest Extraction* – user-side elements that may trigger curiosity; (3) *Intent Inference* – concrete search motivations (find location, similar options, deeper exploration); (4) *Extension & Guidance* – derive the final queries that satisfy immediate needs while enabling further exploration.
Output: Directly output the thinking process, no prefix/suffix; stepwise logical flow; no bare keyword stacking.

Listwise Query Generation Prompt $\mathcal{P}_a$

Role & Task: Generate **5 "extensible" queries** from a user note that guide broader exploration rather than summarizing it.
Workflow: (1) *Analyze* core content, scenario, emotional tone, latent needs; (2) *Diverge* to related activities, categories, crowds, motivations; (3) *Refine* under the extensibility principle.
Constraints:

- **Format:** `<think>[reasoning]</think><answer>[5 queries]</answer>`.
- **Extensibility:** Go beyond literal info (shop/dish names); guide *relevant but distinct* topics.
- **Queries:** exactly 5, each ≤ 15 words, semantically diverse, focused on next-step actions or similar alternatives.
- **Thought:** 50–500 words; the 5 queries must be explicitly derived within it.

Input: Title: {{title}}; Content: {{content}}.

Correction Prompt $\mathcal{P}_c$

Task: Given an Instruction and an Original Response, rewrite the response to improve quality and compliance.
Requirements: (1) *Diversity* – each query differs in semantics and guidance dimension; (2) *Minimum modification* – preserve as many original queries as possible; modify **at most 2**; (3) *Logical consistency* – rewritten CoT must strictly align with the final list; (4) *Non-meta analysis* – the CoT should deeply analyze the original note; *do not* mention "rewrite"/"correct"/"adjust".
Output: Directly output the complete rewritten answer with `<think>` and `<answer>` tags; no additional explanation.
Input: Instruction: {{instruction}}; Original Response: {{response}}.

F Case Study

To concretely illustrate the qualitative difference between ReList and the pointwise DPO baseline, we select two representative cases from the test set, covering accommodation exploration and local-information discovery. For each case, we compare the generated lists along three dimensions: diversity, contextual relevance, and CTR score. In the DPO-baseline column, we underline queries that are semantically near-duplicates of one another.

Case 1: Private Hot-Spring B&B on Chongming Island, Shanghai (accommodation exploration).

NOTE

"Heads up — this is NOT Kyoto, it's Shanghai!!! ✓ pet-friendly ✓ standalone courtyard ✓ private hot spring. Perfect for a weekend of hot-spring soaking, photos, and doing nothing. The B&B is huge, with an outdoor pool, a bar, and more. We picked the courtyard private-onsen tatami room. . . The in-house Japanese restaurant serves incredibly fresh food. Pet-friendliness is outstanding."

Ground-truth user query: *Shanghai Bowu / Namiya Japanese-style B&B*. As shown in Table 6, the DPO baseline mechanically combines "private hot-spring courtyard" with location prefixes, yielding a diversity of only 0.140 and missing salient aspects (Japanese styling, pet-friendliness, upscale tier). ReList instead diversifies along five distinct axes, yielding a **3.7**× improvement in diversity and a **+3.25** absolute CTR gain (4.39 → 7.64).

Case 2: Lize Night Market, Beijing (local-information exploration).

NOTE

"Hot and humid today, went out for an evening stroll. Walked from Lize Tianjie to Lize Weird Market. It has that internet-famous night-market feel, nicely decorated and eye-catching. It's pretty big inside, with trinkets, street food, ring-toss games, and even a beer garden. Lots of stalls, plenty of seating, and it's all surprisingly clean and down-to-earth."

Ground-truth user query: *where is Lize Night Market* (intent: fact-finding). As shown in Table 7, the DPO baseline collapses onto minor perturbations of "Lize Weird Market business hours", giving a diversity of only **0.086** — a direct symptom of independent per-query decoding. ReList decomposes the scene into orthogonal directions (location, cost, browsing, surroundings), delivering nearly **5**× the diversity and a **+1.86** absolute CTR gain (2.39 → 4.25).

Table 6: Case 1: query generation comparison for the Chongming Island private hot-spring B&B note.

#	ReList (Ours)	DPO Baseline
1	Chongming Island private-spring courtyard B&B	Shanghai private hot-spring courtyard
2	Shanghai next-level B&B hotel	Chongming Island private hot-spring courtyard
3	Private hot-spring hotel address	Shanghai Chongming private hot-spring courtyard
4	Courtyard forest-bath spring resort	Chongming Island private-spring courtyard
5	Shanghai next-day B&B Chongming branch price	Shanghai private hot-spring courtyard private B&B
Diversity (r_{div})	**0.516**	0.140
Contextual relevance	0.913	0.960
CTR score	**7.64**	4.39

Table 7: Case 2: query generation comparison for the Beijing Lize Night Market note.

#	ReList (Ours)	DPO Baseline
1	Lize Weird Market address	Lize Weird Market night-market business hours
2	Beijing Lize Tianjie night-market prices	Lize Weird Market business hours
3	Lize Weird Market detailed address	Beijing Lize Weird Market business hours
4	Food and fun around Lize Bridge	Lize Weird Market business hours and location
5	Beijing night-market route map	Lize Weird Market business hours and prices
Diversity (r_{div})	**0.428**	0.086
Contextual relevance	0.867	0.891
CTR score	**4.25**	2.39