Consistency-Aware Online Multi-Objective Alignment for Related Search Query Generation

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

Abstract

Modern digital platforms rely on related search query recommendations to enhance engagement, yet existing methods fail to reconcile click-through rate (CTR) optimization with topic expansion. We propose **CMAQ**, a **C**onsistent **M**ulti-Objective **A**ligned **Q**uery generation framework that harmonizes these goals through three components: (1) reward modeling to quantify objectives, (2) style alignment for format compliance, and (3) consistency-aware optimization to coordinate joint improvements. CMAQ employs adaptive β-scaled DPO with geometric mean rewards, balancing CTR and expansion while mitigating objective conflicts. Extensive offline and online evaluations in a large-scale industrial setting demonstrate CMAQ's superiority, achieving significant CTR gains (+2.3%) and higher human-rated query quality compared to state-of-the-art methods. Our approach enables high-quality query generation while sustaining user engagement and platform ecosystem health.

1 Introduction

Modern digital platforms use related search query recommendation to enhance user experience. An example is illustrated in Figure 1. When users interact with content, the system displays a single related query below it, minimizing disruption. This design serves three key functions: (1) proactive discovery, reducing exploration friction via contextual suggestions; (2) interest scaffolding, enabling gradual topic expansion while avoiding choice overload; and (3) feedback enrichment, where user interactions refine search ranking and content recommendations. By improving user satisfaction and understanding of emerging topics, this mechanism boosts user retention and ecosystem health.

Despite its industrial significance, academic research on related search query recommendation remains limited. Existing methods fall into two categories: retrieval-based and generation-based approaches. Retrieval-based methods (Wang et al., 2023c; Huang et al., 2018; Cao et al., 2008) rely on historical user behavior to retrieve queries from a pool, aligning with sequential patterns but struggling with cold-start content and seamless integration with primary content. In contrast, generation-based methods (Sannigrahi et al., 2024; Wang et al., 2024b), which directly generate queries by considering user interests and context, exhibit superior generalization for cold-start scenarios. Thus, we focus on the generative approach.

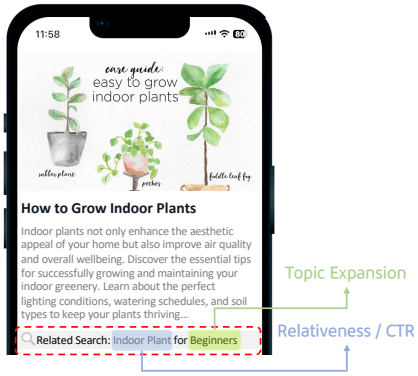


Figure 1: An illustration of the related search query recommendation scenario. A good query should excel in both CTR and topic expansion.

An effective query recommendation system must balance two key dimensions: *relevance* to the user's immediate interests, measurable via click-through rate (CTR), and *topic expansion*, crucial for avoiding filter bubbles (Gao et al., 2023a,b; Bi et al., 2024) and maintaining diversity (Gao et al., 2025b; Kang et al., 2025). However, these objectives often conflict: over-prioritizing relevance leads to narrow recommendations, while excessive focus on topic expansion risks deviating from user intent. Existing methods fail to address this trade-off, motivating our work to align both objectives consistently.

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

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1365–1377
July 28-30, 2025

We leverage large language models (LLMs) (Li et al., 2024; Wang et al., 2023b), whose powerful capabilities make them well-suited for query generation. To mitigate LLM inference latency, we precompute query candidates offline for use in online scenarios. However, directly deploying pre-trained LLMs yields suboptimal performance due to misalignment with task-specific preferences—relevance and topic expansion. Aligning LLMs with these objectives is challenging, as reliable reward signals are hard to obtain: CTR requires extensive online exposure, and topic expansion relies on costly manual annotations. How to consistently enhance the model to achieve both objectives is also critical in this task, *i.e.,* generating queries that offer substantial topic expansion while maintaining a high CTR.

To address these challenges, we propose **C**onsistent **M**ulti-Objective **A**ligned **Q**uery Generation (CMAQ). CMAQ consists of three steps: (1) *precise reward modeling*, training reward models using annotated content-query pairs; (2) *query style alignment*, fine-tuning the LLM to produce correctly formatted queries; and (3) *consistent multi-objective alignment*, introducing a novel training strategy to balance both objectives. The optimization process follows an iterative online DPO paradigm, where generated queries are evaluated by reward models and used to refine the policy. Extensive evaluations demonstrate CMAQ's effectiveness in generating high-quality search queries.

Our key contributions are:

- Formulating related search query recommendation as a multi-objective query generation task.
- Proposing CMAQ, a framework for consistent multi-objective alignment in LLMs, balancing CTR and topic expansion.
- Demonstrating significant improvements via comprehensive offline and online evaluations in a large-scale industrial setting.

2 Related Work

Query Generation. Query generation in content platform is the process of generating new search queries that align with a user's current interests (Li et al., 2024). Existing techniques primarily address scenarios where users have already entered a query prefix, aiming to refine these queries through methods such as query suggestion (Wang et al., 2020; Bacciu et al., 2024), query rewrite (Wang et al., 2023a; Feng et al., 2024; Peng et al., 2024), and personalized query suggestion (Baek et al., 2024; Zhong et al., 2020) incorporating user history and interactions. These approaches assume that users have already demonstrated active search behavior and have initiated a search process.

Our work differs by aiming to provide potential search options to users while they are browsing content, thereby stimulating their interest in active exploration. In this context, early studies on seq2seq models were proposed by (Nogueira et al., 2019; Penha et al., 2023). Recently, some researchers have explored using LLM prompts to generate search terms from context (Sannigrahi et al., 2024), while others have focused on generating search queries in a multimodal context (Wang et al., 2024b). However, these methods overlook the multi-objective alignment problem in query generation. Our approach addresses this gap by simultaneously consider both CTR objective and expansion objective.

Direct Preference Optimization. Learning from human feedback is essential for aligning LLMs with human values (Bai et al., 2022; Ouyang et al., 2022; Ziegler et al., 2019). Recently, DPO-based methods (Rafailov et al., 2023; Ethayarajh et al., 2024; Meng et al., 2024; Wu et al., 2024; Gao et al., 2025a) directly align LLMs with an offline preference dataset, showcasing enhanced training stability and reduced training cost in comparison to traditional RL-based methods (Schulman et al., 2017). Online DPO (Yuan et al., 2024; Xiong et al., 2024; Pang et al., 2024) extends fixed offline preference dataset by continuously updating model preferences from real-time generated responses, enabling dynamic adaptation. Multi-objective DPO (Ramé et al., 2023; Wang et al., 2024a; Zhou et al., 2024; Shi et al., 2024) incorporates multiple criteria for alignment, allowing the model to balance and optimize different human values simultaneously. In industrial scenarios, aligning human preference also attracted attentions, such as query rewrite (Peng et al., 2024), advertising image generation (Chen et al., 2025) and advertising text generation (Wei et al., 2022), however, they primarily focus on aligning their tasks with the CTR objective, overlooking the alignment with broader objectives that impact generation quality, potentially resulting in diminished user experience. In contrast, our method accounts for multi-objective alignment.

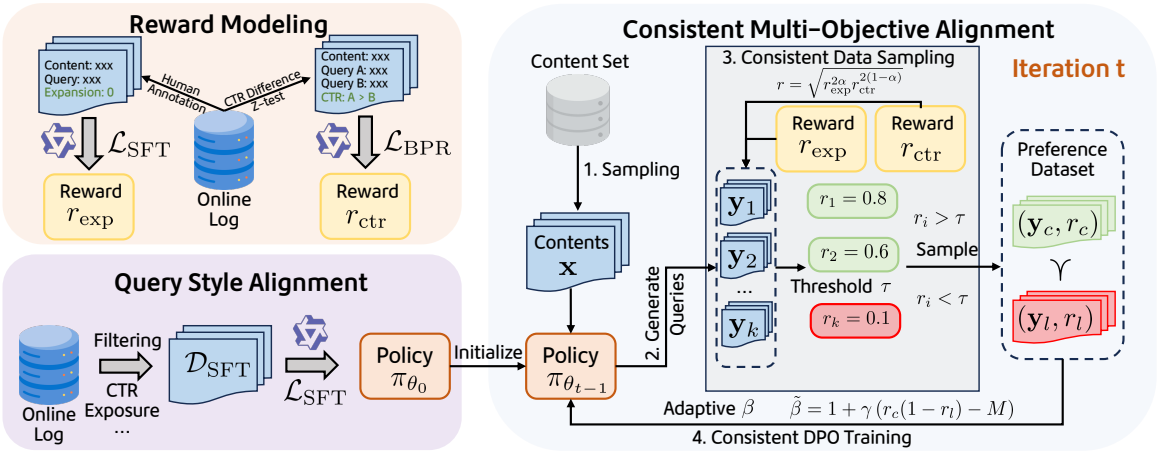


Figure 2: The framework of our proposed CMAQ framework.

3 Methodology

In this section, we introduce our CMAQ framework (*cf.* Figure 2), which consists of three components: reward modeling, query style alignment, and consistent multi-objective alignment. For multi-objective alignment, we primarily focus on the CTR objective and the expansion objective, while our framework is flexible and can be extended to accommodate additional objectives.

3.1 Reward Modeling

To align generated queries with online user preferences, we train two reward models (RMs) using user feedback data, focusing on CTR and topic expansion. These RMs are integrated into the query generation pipeline to guide optimization. Both RMs are based on Qwen2.5-1.5B (Yang et al., 2025) and fine-tuned using LoRA (Hu et al., 2022).

Reward Model for Topic Expansion This RM is designed to determine whether a query extends the context of a given content item, formulated as a binary classification problem. We utilize 337,291 outsourced labeled samples, split 8:2 for training and testing. Among these, 48.8% are labeled as positive (represented by token "1") and the remainder as negative (represented by token "0"). Let $\mathbf{x}$ denote the content and $\mathbf{y}$ the query. The expansion reward $r_{\text{exp}}(\mathbf{x}, \mathbf{y})$ is computed as: $r_{\text{exp}}(\mathbf{x}, \mathbf{y}) = \frac{p(\text{"1"}|\mathbf{x},\mathbf{y})}{p(\text{"0"}|\mathbf{x},\mathbf{y}) + p(\text{"1"}|\mathbf{x},\mathbf{y})}$, where $p(\text{"1"}|\mathbf{x}, \mathbf{y})$ represents the probability of the RM predicting the positive token "1". We use the standard next-token prediction loss to train this RM. The prompt template used for fine-tuning is detailed in Appendix A.1. The final model achieves a classification accuracy of 72.5%.

Reward model for CTR The RM for CTR is designed to predict which of two queries, given the same content, is expected to achieve a higher CTR. This model extends the base architecture with a regression head. We sampled content-query pairs $(\mathbf{x}, \mathbf{y})$ with more than 100 impressions and performed z-tests on impressions and clicks to identify pairs with statistically significant CTR differences ($p < 0.01$). This process yielded 328,328 $(\mathbf{x}, \mathbf{y}_+, \mathbf{y}_-)$ pairs, where $\mathbf{y}_+$ denotes the query with higher CTR for the content $\mathbf{x}$ and $\mathbf{y}_-$ denotes the query with lower CTR for the content $\mathbf{x}$. For the training of the RM, we use Bayesian Personalized Ranking (BPR) loss (Rendle et al., 2009), ensuring reliable distinctions in CTR:

$$\mathcal{L}_{\text{BPR}} = -\log \sigma \left(r_{\text{ctr}}(\mathbf{x}, \mathbf{y}_+) - r_{\text{ctr}}(\mathbf{x}, \mathbf{y}_-) \right). \quad (1)$$

The dataset was split 8:2 for training and testing, achieving a pair accuracy of 91.9%, which measures whether the query with a higher CTR receives a higher reward. In practice, the regression output directly serves as the CTR reward $r_{\text{ctr}}(\mathbf{x}, \mathbf{y})$.

3.2 Query Style Alignment

Initially, we attempted zero-shot or few-shot prompting without fine-tuning the backbone LLM. However, this approach often produced queries that were either non-compliant with instructions, stylistically mismatched with the platform, or contained hallucinated information. To address this, we focused on aligning the query style of the LLM. We constructed a large-scale offline training set $\mathcal{D}_{\text{SFT}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ containing 1,292,031 samples extracted from online logs, leveraging exposure and CTR data to guide this alignment. Supervised

Fine-Tuning (SFT) was then applied to preliminarily align the LLM with the platform's query style, ensuring that generated queries adhere to the expected format and tone:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{\text{SFT}}} \frac{1}{|\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} \log \pi_\theta(y_i|\mathbf{x}, \mathbf{y}_{<i}), \tag{2}$$

where π_θ denotes the model's predicted probability for token y_i given prior tokens and the content.

3.3 Consistent Multi-Objective Alignment

While query style alignment enables the model to mimic real query styles, it does not guarantee high-quality query generation. High-quality queries should not only attract user clicks (high CTR) but also stimulate new search demands (high topic expansion). Therefore, further alignment of these dual objectives is crucial. To minimize reliance on extensive online logs and manual labeling, we employed an online DPO approach. Additionally, we introduced a consistency-aware strategy to mitigate conflicts between the two objectives during both data sampling and training stages.

3.3.1 Consistent Data Sampling

In each iteration t, we sample N content from the offline dataset $\mathcal{D}_{\text{SFT}}$. For each content $\mathbf{x}$, the model from the previous iteration samples k queries $(\mathbf{y}_1, \ldots \mathbf{y}_k) \sim \pi_{\theta_{t-1}}(\cdot|\mathbf{x})$, each evaluated on both objectives. To ensure the same scaling of both rewards, we normalize r_{ctr} into $[0, 1]$. To ensure consistency across both objectives, we used the geometric weighted average $r(\mathbf{x}, \mathbf{y}_i) = \sqrt{r_{\text{exp}}(\mathbf{x}, \mathbf{y}_i)^{2\alpha} r_{\text{ctr}}(\mathbf{x}, \mathbf{y}_i)^{2(1-\alpha)}}$ as the consistency criterion for the queries. By setting two thresholds τ_1 and τ_2 we sample a positive sample $\mathbf{y}_c$ from those with the reward $r > \tau_1$, and a negative sample $\mathbf{y}_l$ with $r < \tau_2$, forming the preference dataset $\mathcal{D}_t = \{(\mathbf{x}, \mathbf{y}_c, \mathbf{y}_l, r_c, r_l)\}$ for the DPO training in the iteration t.

Remark: We use the geometric average instead of the arithmetic average as the overall reward $r(\mathbf{x}, \mathbf{y}_i)$ since it enforces stricter consistency between the two objectives. As illustrated in Figure 3, when one reward approaches zero, the geometric average collapses toward zero regardless of the other reward, ensuring consistent optimization on both rewards.

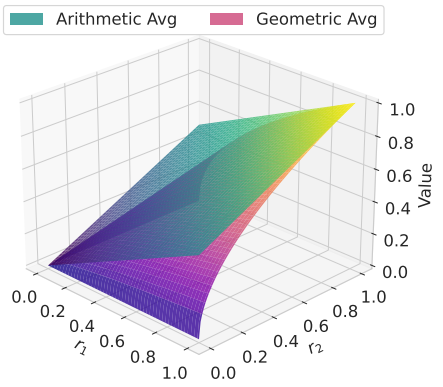


Figure 3: Illustration of the arithmetic average: $(r_1 + r_2)/2$ and geometric average $\sqrt{r_1 r_2}$ over the domain $[0, 1] \times [0, 1]$, demonstrating that the geometric average is more suitable for reflecting consistent multi-objective improvement.

3.3.2 Consistent Training

We adapt and extend DPO (Rafailov et al., 2023) in CMAQ. In DPO, the hyperparameter β controls the strength of KL-divergence regularization between the policy model π_{θ_t} and the reference model $\pi_{\theta_{t-1}}$. The optimal value of β depends on the quality of pairwise preference data (Wu et al., 2024). In our task, the consistency criterion r serves as a proxy for data quality: high-quality pairs exhibit a significantly higher r_c (positive sample) and a substantially lower r_l (negative sample), while low-quality pairs lack this distinction. To account for this variability, we propose a sample-level adaptive β, which dynamically scales β based on the consistency of each training pair. This approach amplifies the influence of high-consistency samples while reducing the impact of low-consistency ones.

For a sample $(\mathbf{x}, \mathbf{y}_c, \mathbf{y}_l, r_c, r_l)$, we compute the sample-level $\tilde{\beta}$ as: $\tilde{\beta} = 1 + \gamma\,(r_c(1 - r_l) - M)$, where $M = \frac{1}{|\mathcal{D}_t|} \sum_{(r_c,r_l)\in\mathcal{D}_t} r_c(1 - r_l)$ represents the average consistency across the dataset. Following (Pang et al., 2024), we incorporate an NLL loss term, weighted by λ, to prevent over-suppression when the chosen query closely resembles the rejected query. The final loss is given by:

$$\mathcal{L}_{\theta_t} = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_c,\mathbf{y}_l)\sim\mathcal{D}_t} \left[\ell(\pi_\theta, \mathbf{x}, \mathbf{y}_c, \mathbf{y}_l) + \lambda \frac{\log \pi_{\theta_t}(\mathbf{y}_c|\mathbf{x})}{|\mathbf{y}_c|} \right],$$
$$\text{with } \ell(\cdot) = \log \sigma \left(\tilde{\beta} \frac{\pi_{\theta_t}(\mathbf{y}_c|\mathbf{x})}{\pi_{\theta_{t-1}}(\mathbf{y}_c|\mathbf{x})} - \tilde{\beta} \frac{\pi_{\theta_t}(\mathbf{y}_l|\mathbf{x})}{\pi_{\theta_{t-1}}(\mathbf{y}_l|\mathbf{x})} \right). \tag{3}$$

4 Experiments

4.1 Experiment Setting

Datasets To the best of our knowledge, no public dataset exists for related search query generation. Therefore, we collected data from a leading content platform. The statistics of the training data are presented in §3. For the test dataset, we randomly sampled 3,124 content items from the training dataset $\mathcal{D}_{\text{SFT}}$. To prevent data leakage, any samples with identical content in the test dataset were excluded from $\mathcal{D}_{\text{SFT}}$. More detailed information on data preprocessing and filtering is provided in A.4.

Baselines We selected two types of comparative approaches. The first type includes non-multi-objective approaches: (1) Zero-shot, where queries are generated directly by LLM without fine-tuning. (2) QSA (Query Style Alignment), as discussed in §3.2, aligns the query style using SFT within $\mathcal{D}_{\text{SFT}}$. (3) DPO (Rafailov et al., 2023), We employ pairwise preference data for CTR reward modeling to fine-tune the QSA model directly using DPO loss.

The second type includes multi-objective alignment approaches, which use the RMs described in §3.1 to obtain two scores for their generated responses, and further fine-tuned on the QSA model: (1) DPO-LW (Zhou et al., 2024), which uses weighted arithmetic average to combines the DPO losses for each objective to form the final loss. (2) DPO-Soup (Ramé et al., 2023), which involves training two models that align with each objective separately, followed by a weighted parameter merge to derive the final model. (3) MORL (Wu et al., 2023), which performs a weighted arithmetic average of the two rewards and then selects the highest and lowest ones to form preference pairs.

Implementation Details All baselines are based on Qwen-2.5-7B-Instruct and fine-tuned using LoRA to ensure a fair comparison. For all DPO-based baselines, we fine-tuned the model for 3 epochs. In the case of multi-objective alignment baselines, the preference dataset is generated at the start of training and remains fixed throughout the training process. For CMAQ, we trained it for 3 iterations, with each iteration comprising 1 epoch. We set the number of training samples per epoch to $N = 20,000$, the number of generated query candidates $k = 8$, the weight for the NLL loss $\lambda = 0.5$, and $\gamma = 0.2$. The trade-off weight in data sampling α is tuned in $[0.2, 0.4, 0.6, 0.8]$ for all multi-objective baselines, larger α indicates more attention on expansion objective. More experimental details can be found in A.4.

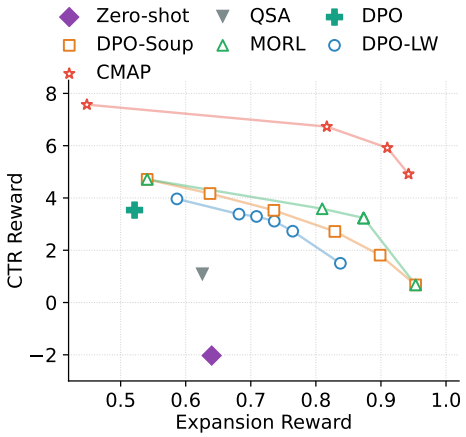


Figure 4: Pareto Fronts of all compared methods.

Evaluation Our framework prioritizes CTR and expansion of generated query: in offline experiments, we directly adopt r_{ctr} and r_{exp} as evaluation metrics, bypassing traditional NLG metrics like BLEU or ROUGE. For online validation, we measure actual CTR on content platforms and incorporate human-annotated quality assessments to holistically evaluate both the practical impact and creative coherence of the outputs.

4.2 Offline Experiments

The performance comparison on the Pareto Fronts of all compared methods is presented in Figure 4. It is important to note that for non-multi-objective baselines, only a single run is conducted as no trade-off is required. From the results, we can observe the following: (1) The Pareto Front of CMAQ significantly exceeds all baseline methods, demonstrating its effectiveness in achieving consistent improvements in both CTR and expansion objectives. (2) Multi-objective methods exhibit superior Pareto Fronts compared to non-multi-objective baselines, highlighting the effectiveness of considering both objectives along with the guidance provided by reward signals. (3) DPO achieves higher CTR rewards while showing a decline in expansion compared to QSA, indicating the presence of conflicts between the two objectives. Therefore, it is crucial to consider consistent optimization for multiple objectives in query generation.

4.3 In-depth Analysis

Ablation Study To validate the effectiveness of each component within our framework, we con-

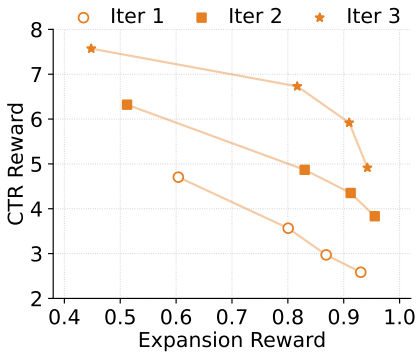


Figure 5: Pareto Fronts of different iterations.

duct ablation studies on three variants of CMAQ: (1) Removing the online query generation at the start of each iteration by utilizing a fixed preference dataset for each iteration, denoted as w/o OT; (2) Removing consistent data sampling by using a weighted arithmetic average instead of a geometric average, denoted as w/o CDS; (3) Removing consistent training by employing a static β in DPO training, denoted as w/o CT.

Table 1 displays the performance of CMAQ and its three variants under two distinct settings, $\alpha = 0.4$ and $\alpha = 0.6$. From the results we can see that (1) removing each component in our framework decreases the performance, validating their effectiveness. (2) The removal of online training leads to a significant deterioration in r_{ctr}, primarily attributed to the absence of iterative on-policy training sample updates. This deficiency substantially diminishes the capacity of training samples to provide effective optimization guidance for model enhancement as the model has already aligned well with the original dataset. (3) The elimination of CDS results in heightened sensitivity to the parameter α, exhibiting a "seesaw effect" where small changes in α lead to sudden shifts in optimization, disproportionately favoring either the CTR or expansion objectives. This issue arises from the limitations of arithmetic mean-based optimization, as discussed in §3, which fails to effectively consistent improvements between dual objectives.

The Impact of Training Iterations To further illustrate the impact of online training, Figure 5 displays the Pareto Front of CMAQ at each iteration. As iterations progress, we observe improved performance, demonstrating the effectiveness of the online training paradigm.

Table 1: Ablation studies on CMAQ. Here, OT, CDS, CT stand for *Online Training*, *Consistent Data Sampling*, and *Consistent Training*, respectively.

Setting	$\alpha = 0.4$		$\alpha = 0.6$	
	r_{ctr}	r_{exp}	r_{ctr}	r_{exp}
CMAQ	6.730	0.817	5.918	0.912
w/o OT	4.055	0.812	3.032	0.906
w/o CDS	6.958	0.481	3.260	0.959
w/o CT	6.672	0.792	5.348	0.910

4.4 Online Experiments

Online Deployment To evaluate the effectiveness of our proposed method in real-world industrial settings, we deployed CMAQ on a local lifestyle information app Dianping, and conducted an online A/B test over a one-week period. We propose to leverage LLMs for query generation as an additional recall pathway in related search scenario. Specifically, we conducted a week-long A/B test involving approximately 3,000,000 contents, where each method employed beam search to sample 5 queries per content. Upon completion of query generation, we further filtered all generated queries through a series of criteria, including lexical quality, relevance, and harmfulness, resulting in the removal of less than 10% of the generated queries. The retained queries were then associated with their respective content and cached in the recall pool. During online service, a fine-grained ranking model determines whether to expose these queries to users. The entire inference process can be executed in offline or nearline modes, allowing for pre-computation and caching of new content, thereby eliminating the need for real-time inference upon user requests and ensuring service efficiency and latency requirements are met.

Online Results The results are presented in Table 2. For data security reasons, CTR results are reported in relative terms, with QSA serving as the baseline model in the A/B test. This experiment gathered over 20 million impressions to ensure the reliability and statistical significance of the CTR results. More detailed online settings can be found in A.4. From the results, we observe the following: (1) DPO demonstrates significant improvement over QSA, highlighting the effectiveness of CTR objective alignment. (2) Multi-objective based methods consistently outperform DPO, suggesting that optimizing for expansion may also contribute positively to CTR. (3) CMAQ achieves the best online CTR

Table 2: The performance of different methods in online A/B test. ΔCTR stands for the relative CTR improvement over QSA: $\frac{\text{CTR}_{\text{method}} - \text{CTR}_{\text{QSA}}}{\text{CTR}_{\text{QSA}}}$.

Method	ΔCTR
DPO	+0.985%
MORL	+1.401%
CMAQ	+2.305%

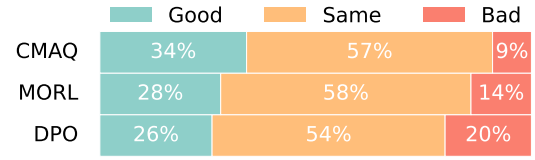


Figure 6: Human Evaluation.

performance, indicating its capability to minimize conflicts between the two objectives.

4.5 Human Evaluation

To validate the quality of queries generated by the model, we conducted a manual GSB (Good-Same-Bad) test on the online methods. Specifically, we randomly selected 200 contents and had human evaluators compare the query quality generated by the online models and QSA. The evaluation criteria included relevance, expansion, and spelling errors. As shown in Figure 6, our proposed CMAQ achieved the best results in comparison with QSA, demonstrating the improvement in query quality offered by our method.

5 Conclusion

In this paper, we introduce CMAQ, a query generation method that formulates related search query generation as a multi-objective alignment task, aligning both CTR and expansion objectives through the online DPO paradigm. We employ consistent data sampling and training strategies to enhance the effectiveness of this multi-objective alignment. Both offline and online experiments demonstrate that CMAQ yields significant improvements in key industrial metrics.

In the future, we aim to take personalization into LLM-based query generation and expand the range of objectives considered in the alignment. We also plan to improve the diversity of the LLM-generated queries while maintaining the performance.

Acknowledgments

This research was supported by Meituan, National Natural Science Foundation of China (62272437,62402470), Anhui Provincial Natural Science Foundation (2408085QF189), and the advanced computing resources provided by the Supercomputing Center of the USTC.

Ethical Considerations

In deploying our query generation model as a supplemental recall mechanism, we prioritize two key ethical principles. (1) **Data Privacy Protection**: All training and inference processes exclusively utilize fully anonymized search session data, with no access to user-specific profiles, search histories, or demographic identifiers. The model operates solely on aggregated query patterns, ensuring complete dissociation from individual users. (2) **Content Safety Risks**: While our framework filters explicit harmful content, automatically generated queries might inadvertently propagate subtle biases from historical search distributions. We mitigate this through regular human audits of sampled outputs and explicit exclusion of sensitive topics during candidate generation.

References

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

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen Herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 3355–3366. ACM.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Shuxian Bi, Wenjie Wang, Hang Pan, Fuli Feng, and Xiangnan He. 2024. Proactive recommendation with

iterative preference guidance. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 871–874. ACM.

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 875–883. ACM.

Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, Yuanjie Shao, Xinge You, Changxin Gao, and Nong Sang. 2025. Ctr-driven advertising image generation with multimodal large language models. *CoRR*, abs/2502.06823.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306.

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.

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. 2025a. Sprec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM Web Conference 2025*, WWW 2025.

Chongming Gao, Mengyao Gao, Chenxiao Fan, Shuai Yuan, Wentao Shi, and Xiangnan He. 2025b. Process-supervised llm recommenders via flow-guided tuning. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, SIGIR 2025.

Chongming Gao, Kexin Huang, Jiawei Chen, Yuan Zhang, Biao Li, Peng Jiang, Shiqi Wang, Zhong Zhang, and Xiangnan He. 2023a. Alleviating matthew effect of offline reinforcement learning in interactive recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2023.

Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023b. Cirs: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems (TOIS)*, 42(1).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhipeng Huang, Bogdan Cautis, Reynold Cheng, Yudian Zheng, Nikos Mamoulis, and Jing Yan. 2018. Entity-based query recommendation for long-tail queries. *ACM Trans. Knowl. Discov. Data*, 12(6):64:1–64:24.

SeongKu Kang, Bowen Jin, Wonbin Kweon, Yu Zhang, Dongha Lee, Jiawei Han, and Hwanjo Yu. 2025. Improving scientific document retrieval with concept coverage-based query set generation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10-14, 2025*, pages 895–904. ACM.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022,*

NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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, Zachary 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. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

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*.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. 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*.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: bayesian personalized ranking from implicit feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 452–461. AUAI Press.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A. Smith, and Simon S. Du. 2024. Decoding-time language model alignment with multiple objectives. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. 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 8642–8655. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

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.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023b. Generative recommendation: Towards next-generation recommender paradigm. *CoRR*, abs/2304.03516.

Yu Wang, Zhengyang Wang, Hengrui Zhang, Qingyu Yin, Xianfeng Tang, Yinghan Wang, Danqing Zhang, Limeng Cui, Monica Xiao Cheng, Bing Yin, Suhang Wang, and Philip S. Yu. 2023c. Exploiting intent evolution in e-commercial query 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 5162–5173. ACM.

Zheng Wang, Bingzheng Gan, and Wei Shi. 2024b. 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.

Penghui Wei, Xuanhua Yang, Shaoguo Liu, Liang Wang, and Bo Zheng. 2022. CREATER: ctr-driven advertising text generation with controlled pre-training and contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2022, Hybrid: Seattle, Washington, USA + Online, July 10-15, 2022*, pages 9–17. Association for Computational Linguistics.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. β-dpo: Direct preference optimization with dynamic β. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. 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*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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. 2025. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10586–10613. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Appendix

A.1 Prompts

Here, we introduce the prompts used in the query generation and expansion reward models. For the CTR reward model, as it is treated as a regression task, we do not design a specific prompt template. Instead, the input to the CTR reward model is simply the concatenation of (title, content, shopinfo, query).

Prompt for Query Generation

You are a user of a leading local lifestyle information platform that provides shop info, consumer reviews, discounts, and nearby lifestyle information. You often browse user-generated content and excel at summarizing and extending related interest queries to help other users explore more related information.
Requirements:
1. Provide only one answer, keep it within 15 words.
2. Output the answer directly, without any explanations or unnecessary prefixes.
3. The answer should be related to the content but not just a summary, guiding users to search for more related topics.
Given a note, please summarize and extend the interest queries for the content.
##Note Content
Title: {{title}}
Content: {{content_body}}
Shop info: {{shopinfo}}
Answer:

Prompt for Expansion Reward Model

You are a search term quality assessment expert. Based on the following note content and query, score the query's expansion (0 or 1), and output the result in the specified format without explanations.
Expansion: Does the search query include information beyond the note content that can spark user interest for further exploration? It might involve novel, interesting, or trending topics that seem worth delving into.
Score 0: Completely redundant information (directly copying POI name/title queries), with no apparent extensibility, as the information is fully covered by the note content, and users can get complete information without further clicking.
Score 1: Has a certain extensibility. Even if the note doesn't mention this information, if the query can guide users to acquire new useful information (like reservation methods) or encourage comprehensive exploration of the place (like "exploring shop" queries), it is considered to have extensibility.
##Note Content
Title: {{title}}
Content: {{content_body}}
Shop info: {{shopinfo}}
Query: {{query}}
Answer:

A.2 The Pseudo Code of Consistent Multi-Objective Alignment

Algorithm 1: Consistent Multi-Objective Alignment

Data: Offline content dataset $\mathcal{D}_{\text{SFT}}$, QSA model $\pi_{\theta_{\text{QSA}}}$, Threshold τ_1, τ_2, Adaptation rate γ, Trade-off parameter α, Sample number N, Generation number k, Max iteration T

Initialize policy $\pi_{\theta_0} \leftarrow \pi_{\theta_{\text{QSA}}}$;

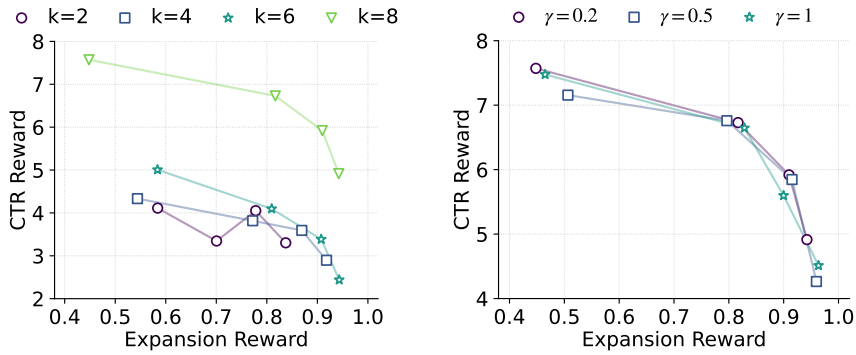
for *iteration* $t = 1, 2, \dots T$ **do**

- $\mathcal{D}_t \leftarrow \emptyset$;
- Sample contents $\{\mathbf{x}\}_1^N \sim \mathcal{D}_{\text{SFT}}$;
- **for** *content* $\mathbf{x} \in \{\mathbf{x}\}_1^N$ **do**
  - Generate queries $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\} \sim \pi_{\theta_{t-1}}(\cdot|\mathbf{x})$;
  - Compute rewards $r_i = \sqrt{r_{\text{exp}}^{2\alpha}(\mathbf{x}, \mathbf{y}_i) r_{\text{ctr}}^{2(1-\alpha)}(\mathbf{x}, \mathbf{y}_i)}$ for each $\mathbf{y}_i$;
  - $\mathcal{D}_{\text{pos}}, \mathcal{D}_{\text{neg}} \leftarrow \emptyset, \emptyset$;
  - **for** *query* $i = 1, 2, \dots k$ **do**
    - **if** $r_i > \tau_1$ **then**
      - $\mathcal{D}_{\text{pos}} \leftarrow \mathcal{D}_{\text{pos}} \cup \{(\mathbf{x}, \mathbf{y}_i, r_i)\}$;
    - **if** $r_i < \tau_2$ **then**
      - $\mathcal{D}_{\text{neg}} \leftarrow \mathcal{D}_{\text{neg}} \cup \{(\mathbf{x}, \mathbf{y}_i, r_i)\}$;
  - **if** $\mathcal{D}_{pos} \neq \emptyset$ *and* $\mathcal{D}_{neg} \neq \emptyset$ **then**
    - $(\mathbf{y}_c, r_c) \sim \mathcal{D}_{\text{pos}}$;
    - $(\mathbf{y}_l, r_l) \sim \mathcal{D}_{\text{neg}}$;
    - $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{(\mathbf{x}, \mathbf{y}_c, \mathbf{y}_l, r_c, r_l)\}$;
- Compute averge reward $M = \frac{1}{|\mathcal{D}_t|} \sum_{(r_c, r_l) \in \mathcal{D}_t} r_c(1 - r_l)$;
- **for** *data sample* $(\mathbf{x}, \mathbf{y}_c, \mathbf{y}_l, r_c, r_l) \in \mathcal{D}_t$ **do**
  - Compute adaptive $\tilde{\beta} = 1 + \gamma\,(r_c(1 - r_l) - M)$;
  - Perform Consistent DPO Training via Equation (3);

A.3 Data Collection

We construct the dataset $\mathcal{D}_{\text{SFT}}$ where each sample $(\mathbf{x}, \mathbf{y})$ is a tuple of (content, query). The construction procedure of $\mathcal{D}_{\text{SFT}}$ mainly includes the following steps:

- **Core Metric Aggregation.** We first aggregate behavioral signals (page views, clicks) at the content-query level through temporal summation, with the time spans one year. This initial phase establishes baseline engagement metrics and computes derived indicators including CTR. A minimum exposure threshold eliminates statistically insignificant observations.
- **Multi-Dimensional Filtering.**The raw dataset undergoes successive quality filters:
  - Lexical constraints: Remove short/non-compliant queries through length thresholds and regex pattern matching.
  - Engagement thresholds: Eliminate low-CTR entries through percentile-based cutoffs
  - Commercial term exclusion: Filter queries containing promotional phrases via predefined blocklists
  - Semantic redundancy checks: Exclude queries exhibiting high similarity to shop names through normalized Levenshtein distance calculations
- **Diversity-Preserving Sampling.** To ensure categorical diversity and prevent domain dominance in the training corpus, we implement a stratified sampling strategy grounded in content taxonomy. The dataset is first partitioned by content categories. Within each categorical partition, entries are ranked



(a) (b)

Figure 7: (a) The Pareto Front of CMAQ under different query sample times $k \in [2, 4, 6, 8]$. (b) The Pareto Front of CMAQ under different scaling coefficient γ in obtaining $\tilde{\beta}$, where $\gamma \in [0.2, 0.5, 1]$.

through a composite scoring metric prioritizing CTR while considering auxiliary quality signals. A maximum cap of 10,000 samples per category is enforced to prevent the bias of prevalent domains.

Finally, we collected $\mathcal{D}_{\text{SFT}}$ for both quality style alignment and consistent multi-objective alignment processes. The size of $\mathcal{D}_{\text{SFT}}$ is 1,292,031.

A.4 Detailed Experiment Settings

For all fine-tuning experiments in each iteration, we utilize PyTorch 2.1.0[1] (Paszke et al., 2019) in conjunction with HuggingFace's TRL framework[2]. Experiments are executed on eight A100 GPUs, with each iteration requiring approximately 10 GPU hours, including query generation, rewarding and training. We employ the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 and a cosine learning rate schedule incorporating 20 warmup steps. The temperature is set to 1.5 in generation k queries to ensure the diversity for iterative DPO training. The training process spans 1 epochs with a global batch size of 32. For LoRA training, we set the rank $r = 32$, and the $\alpha = 8$. For online inference, we utilize vLLM[3] (Kwon et al., 2023) for speed-up.

A.5 Supplementary Experimental Results

We conducted additional experiments to investigate the impact of the sampling number k and the scaling coefficient γ in Equation (3) on the performance.

The impact of sample times k Figure 7a illustrates that the performance of CMAQ improves as k increases, suggesting that additional sampling instances contribute to more diverse information during training. As the number of sample times rises with k, we select $k = 8$ for our final model, balancing the trade-off between performance and efficiency.

Parameter sensitivity of γ Figure 7b indicates that CMAQ exhibits robustness across various values of γ. This suggests that the method maintains its effectiveness despite changes in the hyperparameter settings, making it adaptable to different conditions.

[1] https://pytorch.org/
[2] https://github.com/huggingface/trl
[3] https://github.com/vllm-project/vllm