

WeFair: Fidelity-Aware Alignment from Human Feedback for Personalized Query Rewriting in Short-Video Search

Yingtao Ren*
Aolin Li*
Cheng Cheng*
yingtaoren@tencent.com
churenli@tencent.com
trcheng@tencent.com
Weixin Group, Tencent
China

Huiyun Hu
Huasheng Liang
Baoping Xie
huiyunhu@tencent.com
watsonliang@tencent.com
baopingxie@tencent.com
Weixin Group, Tencent
China

Jun Yuan
Haijun Wu
Chenxing Wang^{†*}
junyuan@tencent.com
ethanhjwu@tencent.com
magnuswang@tencent.com
Weixin Group, Tencent
China

Abstract

Short-video search requires compact query rewriters that infer personalized intent from brief, ambiguous queries under strict latency constraints. The challenge is to decide when personalization is warranted and generate a reformulation that preserves the user’s immediate information need. A rewriter optimized for fluency or engagement may still introduce unsupported details and redirect the search toward a different intent. Existing methods often under-utilize users’ actual reformulations and behavioral traces, relying instead on synthetic reasoning and indirect behavioral rewards that may produce more engaging candidates but cause intent drift. To build an effective rewriter that remains faithful to user intent, we propose WeFair, Fidelity-Aware Alignment from Human Feedback for Personalized Query Rewriting in WeChat short-video search. WeFair assigns real user behavioral signals to a progression from demonstration to preference to verification. We first mine actual human reformulations from search sessions, then apply multidimensional LLM screening and contextual verification to filter unsupported transitions without extensive manual annotation. During training, we begin with a supervised curriculum that progressively builds rewriting capabilities from large-scale human personalized query demonstrations, and then use direct preference optimization to align with users’ subsequent consumption preferences. In the final on-policy optimization stage, we construct a strict verification reward based on actual human reformulations to further boost performance with group relative policy optimization. Further analysis shows that auxiliary rewards can activate more rollout groups in on-policy optimization but reduce intent fidelity. Online A/B testing in WeChat short-video search shows a 1.576% relative increase in effective clicked-query rate (VV> 10s) and a 1.731% relative reduction in query reformulation rate.

Keywords

Personalized query rewriting, short-video search

1 Introduction

Personalized query rewriting is essential when a brief search query does not fully express the user’s information need. In short-video search, the intended meaning may depend on recent searches, viewed content, or other context available when the query is issued.

*These authors contributed equally to this work.

[†]Corresponding author.



Figure 1: Overview of intent drift in WeChat short video search. The original query “crepe cake” is redirected toward recipes instead of the intended eating videos. WeFair uses human feedback through demonstration, preference, and verification to mitigate intent drift.

Rewriting can expose this information to retrieval by producing a more informative expression, but it must also satisfy strict latency constraints [3, 4]. Therefore, the practical objective is training a compact rewriter model that can resolve ambiguity without replacing the user’s current intent and needs. Furthermore, the rewriter model also needs to accurately decide when to rewrite or preserve the original query.

Although general-purpose large language models (LLMs) exhibit strong reasoning capabilities, they fail to accurately understand diverse user intents and require several seconds per inference, which is unacceptable for real-time search. To address these limitations, some approaches use synthetic Chain-of-Thought traces to distill reasoning capabilities from large teacher LLMs into a smaller

rewrite model, and then use large-scale poster consumption feedback to align them with platform-specific user preferences.

However, consumption-oriented alignment does not necessarily preserve the user’s immediate information need. When insufficiently constrained by direct evidence of user intent, a rewriter may favor queries that retrieve engaging content while redirecting the original search, resulting in intent drift. To better capture personalized intent, recent methods incorporate actual user reformulations, which provide direct evidence of how users clarify their information needs. PicQue uses historical reformulations as context for generating training examples that subsequently undergo manual verification, while WeWrite directly uses mined reformulations for supervised fine-tuning and reward construction in GRPO. Despite this progress, how to exploit the distinct roles of real user signals across training stages remains underexplored. While these advances demonstrate the value of real user signals, reformulations and subsequent consumption feedback differ in how directly they reflect personalized intent, and their fine-grained use across training stages remains underexplored.

To bridge the gap, we propose WeFair, a fidelity-aware alignment framework for training compact personalized query rewriters in WeChat short-video search. Its central idea is to match supervision fidelity to the requirements of each training stage, improving how human feedback guides a small model toward faithful and efficient rewriting. As illustrated in Figure 1, this progression can enable a compact model to acquire personalized query rewriting capabilities without intent drift.

WeFair combines automated data construction with three complementary training stages. We mine actual reformulations from search sessions and apply LLM screening and contextual verification to retain demonstrations grounded in user intent. Training then proceeds through three stages, each with a distinct purpose. First, a supervised fine-tuning (SFT) curriculum progresses from session-to-query prediction to context-conditioned reasoning and finally direct query rewriting, building intent understanding before consolidating rewrite-or-preserve behavior. The broad coverage of human demonstrations prepares the compact policy model to generate useful candidates for later optimization. Second, direct preference optimization (DPO) learns from candidate pairs distinguished by subsequent consumption outcomes, increasing the probability of rewrites associated with better user experiences. This stage uses consumption feedback as comparative evidence to refine candidate preferences. Finally, group relative policy optimization (GRPO) applies a strict verification reward based on exact match with actual human reformulations or preservation. This final stage anchors policy updates to observed human expressions without adding graded similarity, retrieval, or other judge rewards. At inference time, the rewrite model directly outputs a short rewrite or a preservation decision, avoiding explicit CoT generation. Figure 2 shows the overview of WeFair.

Our further analysis suggests that the three training stages differ in their tolerance for supervision with varying intent fidelity. SFT can tolerate lower-fidelity supervision to acquire broad rewriting capabilities, while DPO can incorporate preferences inferred from subsequent consumption to refine candidate selection. In contrast, final GRPO optimization requires strict fidelity to the intended target: auxiliary rewards can activate more rollout groups while

introducing intent drift. Experiments demonstrate WeFair’s effectiveness in personalized rewriting. Overall, our main contributions can be summarized as follows:

- (1) We examine the role of supervision fidelity in personalized query rewriting, showing how consumption-based optimization can induce intent drift.
- (2) We propose WeFair, a fidelity-aware alignment framework that matches real user feedback to different training stages, combining broad capability acquisition, preference learning, and strict verification to mitigate intent drift.
- (3) We evaluate WeFair through online A/B testing to assess rewriting effectiveness.

2 Related Work

2.1 Personalized Query Rewriting

LLM-based query rewriting improves retrieval by enriching query representations and adapting them to search domains. Query2doc generates pseudo-documents, while CardRewriter incorporates platform knowledge through knowledge cards [4, 22]. Domain adaptation further aligns query expressions with retrieval requirements. BEQUE combines supervised rewriting with retrieval feedback [17], and negation-aware rewriting preserves explicit constraints [6]. MAAQR and QP-OneModel extend this direction through adaptive refinement and unified query understanding [8, 26]. Beyond representation enrichment, IterQR and Think-Then-Rewrite incorporate reasoning into retrieval-oriented training [1, 11], while MiniELM combines lightweight distillation with simulated feedback [14]. For personalized search, these capabilities additionally require evidence of what the current user intends to find, which utilizes user context and observed behavior to resolve ambiguity and determine when rewrite is essential. PicQue combines hybrid reasoning supervision with GRPO to address over-rewriting without online CoT generation [3], while Think-to-Personalize uses purchase history to support personalized retrieval [9]. WeWrite mines actual user reformulations and trains with SFT and GRPO using lexical similarity and conditional query-volume rewards [2]. Together, these studies establish complementary roles for contextual evidence and users’ expressed reformulations in understanding personalized intent.

Behavioral feedback also raises a distinction between search utility and fidelity to an issued query. OneBar emphasizes content-grounded recommendations, while short-video and conversational query suggestion address consumption utility and intent coverage [13, 20, 21]. DVCQR aligns separate rewrites with sparse and dense retrieval [12]. These objectives can improve discovery, but their gains alone do not establish intent preservation.

2.2 LLM Post-Training and Feedback Alignment

Post-training methods provide different mechanisms for learning from demonstrations, preferences, and rewards. SFT learns demonstrated responses [16], whereas DPO optimizes chosen–rejected preferences without separately training a reward model [18]. PPO uses clipped policy updates [19], and GRPO estimates advantages within sampled groups without a learned value network [5]. Beyond algorithm selection, recent work examines which supervision is useful at different learning stages. Stage-specific data selection

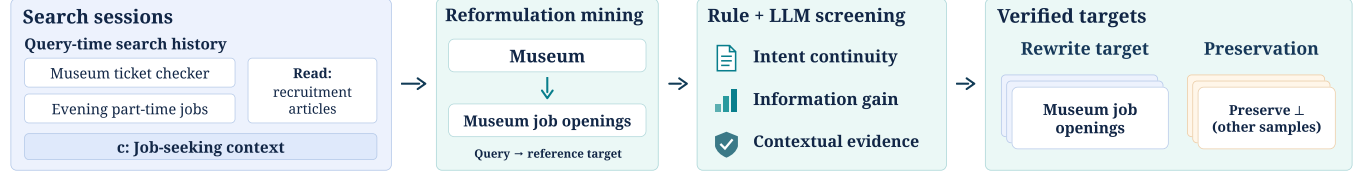
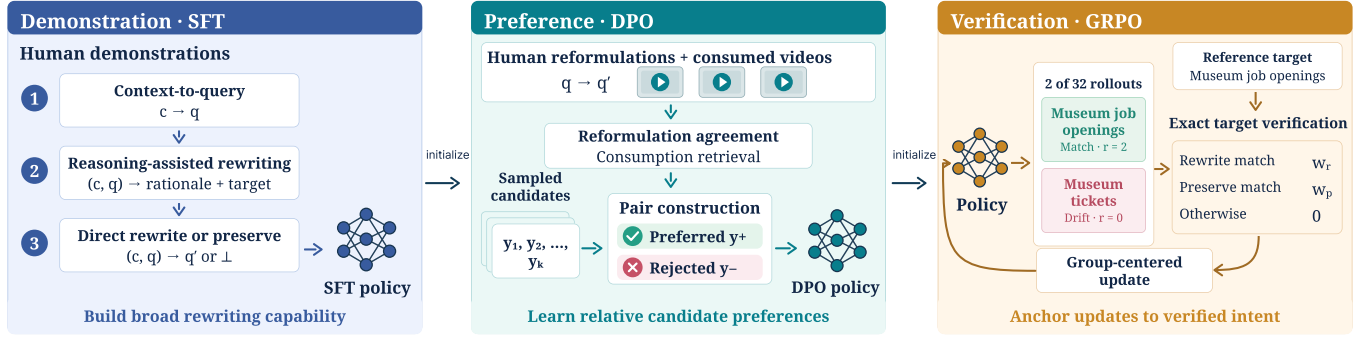
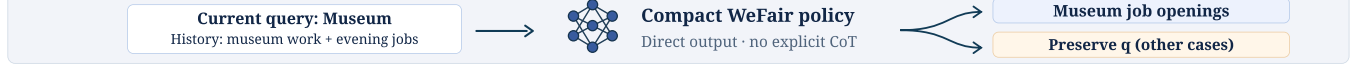
(a) Human feedback construction**(b) Fidelity-aware progressive alignment****(c) Online inference**

Figure 2: Overview of WeFair: (a) Search sessions yield screened reformulation and preservation targets. (b) Curriculum SFT builds rewriting capabilities, DPO learns preferences from reformulations and consumption feedback, and GRPO applies strict target verification to constrain intent drift. (c) The compact policy directly rewrites or preserves the query using query-time context, without explicit chain-of-thought generation.

distinguishes capability acquisition from reinforcement [7]. SmartSearch uses a process-reward-guided curriculum [23]. Hindsight self-distillation further derives corrections and preferences from follow-up user interactions [10]. This work motivates matching supervision to both the policy’s capabilities and the evidence available from users. In search applications, this allocation also depends on what feedback measures. AWHCP combines SFT, DPO, and GRPO with quality, format, and recall rewards [15], while CoRe uses deployed relevance feedback [24]. Such signals align generation with search outcomes, but provide different evidence about intended reformulations. Comparisons under compute constraints further show that on-policy rewriting need not outperform prompted alternatives [25], motivating examination of the optimization signals alongside the algorithms.

Reward-based optimization consequently requires considering the reliability of candidate-level feedback. TTRL learns from majority-voted pseudo-labels and shows that reward accuracy can remain high despite inaccurate labels: an incorrect label can still assign correct negative rewards to other incorrect answers [27]. This separates source-label accuracy from the reliability of induced rewards.

3 Methodology

3.1 Task Formulation and Framework Overview

Task Formulation. In short-video search system, given a query q and user context c available at query time, a policy $\pi_\theta(y | x)$, where $x = (c, q)$, generates a reformulation $\hat{q}$ or a preservation marker $\perp$.

The reformulation serves as an additional retrieval query, and the preservation marker ensures that only the original query is retained. We define intent drift as changing or omitting information required by the user, or introducing personalized constraints unsupported by x . The objective is to generate context-supported reformulations while preserving the original query.

Framework Overview WeFair organizes training according to intent fidelity: how reliably supervision reflects the user’s immediate information need. Human reformulations provide direct but incomplete evidence of intent, whereas subsequent consumption offers indirect preference signals. Accordingly, curriculum SFT builds rewriting capabilities from broad demonstrations, DPO learns relative preferences from consumption-grounded pairs, and GRPO anchors final refinement through strict agreement with screened reformulation and preservation targets.

3.2 Human Feedback Construction

WeFair utilize context-to-query records $C = \{(c, q)\}$ and user reformulations for demonstrations, preference learning, and verification. We first extract context-to-query records by pairing user-entered queries with preceding context and requiring at least one retrieved video to be watched for more than ten seconds. For Reformulation candidates $\mathcal{R}$, we link eligible adjacent queries (q, q') within the same session to the context preceding q . Both sources exclude queries issued through system suggestions. Session adjacency alone does not establish intent continuity. Therefore, we remove tagged

Algorithm 1 Human feedback construction**Require:** Candidates $\mathcal{R}$, natural pool $\mathcal{N}$, parameters Θ , exclusions $\mathcal{X}$ **Ensure:** Reformulation pool $\mathcal{P}$, preservation pool $\mathcal{B}$

```

1:  $\mathcal{P} \leftarrow \emptyset; \mathcal{H} \leftarrow \emptyset$ 
2: for each valid  $(q, q', c) \in \mathcal{R}$  do
3:    $(c_s, e) \leftarrow \text{PREPAREEVIDENCE}(c)$ 
4:   if Rules 1–4 reject under  $\Theta$  then continue
5:    $d \leftarrow \text{SCREENCANDIDATE}(q, q', c_s)$ 
6:   if  $d = 0$  then
7:      $\mathcal{H} \leftarrow \mathcal{H} \cup \{(q, q', c)\}$ 
8:   else if  $d = 1$  then
9:      $t \leftarrow \text{PRESERVATIONGATE}(q, q', c_s)$ 
10:    if  $t = 0$  then  $\mathcal{P} \leftarrow \mathcal{P} \cup \{(c, q), (q', c)\}$ 
11:  end if
12: end for
13:  $\mathcal{B} \leftarrow \text{SELECTHARD}(\mathcal{H}) \cup \text{SELECTNATURAL}(\mathcal{N})$ 
14: return  $\text{CLEANANDEXCLUDECONFLICTS}(\mathcal{P}, \mathcal{B}, \mathcal{X})$ 

```

suggestions and lookup-table fragment replacements from c to obtain screening context c_s , then strip selected headers and history-entry prefixes to derive rule evidence e . These views serve screening and leave the policy input unchanged. Algorithm 1 summarizes the overall construction process.

Hard-rule screening. After getting the initial candidates, we then conduct a rule-based screening. Let $S(t)$ denote a string’s character set, $a = s(q)$, $b = s(q')$, and $U = S(b) \setminus S(a)$, where s removes whitespace, non-word characters, and underscores. Reject a candidate if any condition holds:

- (1) *No character gain:* $U = \emptyset$.
- (2) *Function-word-only gain:* the added characters belong to a fixed Chinese function-word lexicon and introduce a function word.
- (3) *Insufficient overlap:* $J(q, q') = |S(q) \cap S(q')| / |S(q) \cup S(q')| \leq 0.06$, with the threshold calibrated on a small annotated set.
- (4) *Unsupported change:* no contextual match passes the new trigram, bigram, single-character, or common-prefix fallback conditions.

The contextual alternatives use occurrence thresholds in e and are evaluated in that order. Only candidates passing all vetoes reach LLM screening. These lexical tests provide coarse evidence, which requires subsequent semantic validation. Accordingly, only candidates that pass all these hard-rule vetoes proceed to LLM screening, where such semantic validation is carried out.

LLM screening. This step operationalizes the semantic validation by checking three conditions: intent continuity I , information gain G , and contextual bridging B . These conditions are evaluated by Gemini 3.5 Flash:

$$K_{\text{LLM}}(q, q', c_s) = I(q, q') \wedge G(q, q') \wedge B(q, q', c_s).$$

Continuity preserves the subject, task, answer type, and essential constraints, and gain requires useful retrieval information or supported entity recovery. Bridging rejects unsupported evidence (Level 1) and isolated signals such as target replay or broad interests (Level 2). Level 3 requires at least two independent, consistent signals supporting a stable attribute or nearly unique entity recovery; Level 4 requires direct evidence linking the original subject or task to the added detail. Stable disambiguation admits Level 3; relational

requests generally require Level 4. The model returns acceptance $d = 1$ or rejection $d = 0$ with a brief reason. Malformed judgments are discarded. Approximately one third of valid decisions accept the candidate.

Lexical retention and transfer verification. For LLM-accepted candidates, PRESERVATIONGATE retains pairs where the preprocessed original query is a substring, ordered subsequence, or character multiset contained in the target. These lexical shortcuts preserve characters, not necessarily intent. Other pairs reuse an existing semantic-transfer judgment for the same record or receive a new one; only a valid non-transfer label ($t = 0$) passes. Approximately seven tenths of LLM-accepted candidates enter $\mathcal{P}$, retaining the observed target $y^* = q'$.

Preservation labels. For preserved query candidates, natural examples come from an existing non-reformulation pool $\mathcal{N}$, and the hard examples are selected from valid LLM rejections $\mathcal{H}$. Both receive $((c, q), \perp)$ in $\mathcal{B}$. Rejected candidates and screening explanations are excluded from policy input. These labels prescribe conservative preservation without establishing that every possible rewrite is invalid. Rule failures, malformed judgments, and semantic-transfer rejections are not automatically converted into preservation examples.

Finally, all candidates undergo validity checks, evaluation exclusion, and label-conflict removal before sample records are assigned to mutually exclusive training stages. This partitioning is at the record level and does not imply request-level or user-level separation.

3.3 Fidelity-Aware Progressive Alignment

The training of WeFair follows demonstration, preference, and verification, with each stage initializing the next.

Demonstration: curriculum SFT. The curriculum consists of three phases. Phase 1 trains the model to predict the query q from preceding context c . Phase 2 trains it to learn context-conditioned rationales that explain either a reformulation or a preservation decision. Phase 3 trains it to predict the final target directly from $x = (c, q)$. Context-to-query records provide broad behavioral coverage, while the smaller adaptation subset teaches evidence-supported rewriting and the serving format. The rationales are auxiliary explanations generated by Gemini-3.1-Pro-Preview, not observed human reasoning. For phase $k \in \{1, 2, 3\}$ and input–target sequence pair (u, v) , the objective function is:

$$\mathcal{L}_{\text{SFT}}^{(k)} = -\mathbb{E}_{(u,v) \sim \mathcal{D}_S^{(k)}} \sum_{t=1}^{|v|} \log \pi_{\theta}(v_t \mid u, v_{<t}). \quad (1)$$

Each phase continues from the preceding checkpoint; after the final direct-prediction phase, we obtain π_{SFT} .

Preference: DPO with zero-consumption negatives. For each preference input, the policy generates 64 candidates $\mathcal{Y}(x)$, and we construct up to two DPO pairs from the same input. Let $r_i(x, y)$ be the rank of the i -th subsequent relevant consumed video in candidate y ’s retrieval results. Using up to three such videos as user preference, a chosen candidate y^+ must have at least one relevant consumed video appear in its top-20 retrieval results. A reject candidate y^- satisfies no relevant consumed video in its top-100 retrieval results. The first pair uses the human reformulation q' as the positive,

yielding (x, q', y_1^-) . The second pair uses a consumption-supported chosen candidate, yielding $(x, y_{\text{cons}}^+, y_2^-)$. If only one valid negative candidate is available, we set $y_1^- = y_2^-$ and share it across both pairs. For preserved samples, we simply use the preservation marker $\perp$ and any generated candidate y^- , yielding $(x, \perp, y^-)$. DPO increases the relative likelihood of human reformulations or consumption-supported candidates over observed zero-consumption alternatives. Consumption thus guides comparative preference learning with relatively high fidelity. Because engagement can differ from the immediate information need, final alignment returns to explicit reformulation and preservation markers.

Verification: human-grounded GRPO. Verification is the final mechanism for constraining optimization to direct target evidence with GRPO. Starting from the DPO policy, we sample 32 responses for each $(x, y^*) \in \mathcal{D}_V$. After deterministic parsing, a valid response y receives

$$R(y, y^*) = \begin{cases} w_r, & y = y^*, \quad y^* \neq \perp, \\ w_p, & y = y^* = \perp, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $w_r, w_p > 0$ weight exact matching with the retained human query q' and preservation, and any other responses receive zero reward, ensuring the highest fidelity signal supervision. Our further analysis shows that the high-fidelity reward function outperforms other designs with graded similarity, retrieval, or learned-judge rewards. This conservative final stage anchors updates to human reformulations and preservation labels, mitigating consumption-driven intent drift.

4 Experiment

4.1 Online Evaluation

We evaluate WeFair through online A/B testing in WeChat short-video search. Relative to the control, the deployed model improves the effective clicked-query rate (VV> 10s) by 1.576% and reduces query reformulation rate (QRR) by 1.731%. Here, VV> 10s measures the fraction of exposed queries with at least one video consumption exceeding ten seconds. These results indicate improved effective consumption and fewer subsequent query revisions.

5 Conclusion

We presented WeFair, a fidelity-aware alignment framework for compact personalized query rewriting in short-video search. WeFair addresses intent drift by assigning human feedback to complementary training roles: curriculum SFT builds rewriting capabilities, DPO learns relative preferences, and GRPO anchors final refinement to strict agreement with human reformulations and preservation targets. The resulting policy directly rewrites or preserves the query without generating reasoning traces. Online A/B testing shows a 1.576% relative increase in effective clicked-query rate and a 1.731% relative reduction in query reformulation rate, supporting the framework's practical value for personalized search.

References

- [1] Shangyu Chen, Xinyu Jia, Yingfei Zhang, Shuai Zhang, Xiang Li, and Wei Lin. 2025. IterQR: An Iterative Framework for LLM-based Query Rewrite in e-Commercial Search System. *arXiv preprint arXiv:2504.05309* (2025). arXiv:2504.05309 <https://arxiv.org/abs/2504.05309>
- [2] Cheng Cheng, Chenxing Wang, Aolin Li, Haijun Wu, Huiyun Hu, Juyuan Wang, and Dongliang Liao. 2026. When & How to Write for Personalized Demand-aware Query Rewriting in Video Search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 4528–4533. doi:10.1145/3805712.3808395
- [3] Xu Chu, Angela Li, Jiaming Zhang, Wei Li, Zhijie Tan, Dawei Yin, Shuaiqiang Wang, and Daiting Shi. 2026. Accurate and Efficient Personalized Query Rewriting in Baidu Search. In *Proceedings of the ACM Web Conference 2026*. ACM, 7902–7913. doi:10.1145/3774904.3792815
- [4] Peiyuan Gong, Feiran Zhu, Yaqi Yin, Chenglei Dai, Chao Zhang, Kai Zheng, Wentian Bao, Jiaxin Mao, and Yi Zhang. 2026. CardRewriter: Leveraging Knowledge Cards for Long-Tail Query Rewriting on Short-Video Platforms. In *Proceedings of the ACM Web Conference 2026*. ACM, 8029–8040. doi:10.1145/3774904.3792809
- [5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirog Ma, Xiao Bi, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (2025), 633–638. <https://www.nature.com/articles/s41586-025-09422-z>
- [6] Mengtian Guo, Mutasem Al-Darabsah, Choon Hui Teo, Jonathan May, Tarun Agarwal, and Rahul Bhagat. 2025. Learning to Rewrite Negation Queries in Product Search. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. Association for Computational Linguistics, Abu Dhabi, UAE, 575–582. <https://aclanthology.org/2025.coling-industry.49/>
- [7] Chongyang He, Rui Zhang, Zixuan Wang, and Xin Li. 2026. Learning What to Learn: Stage-Specific Data Sets for SFT-then-RL in Small Language Model Reasoning. *arXiv preprint arXiv:2606.04466* (2026). arXiv:2606.04466 <https://arxiv.org/abs/2606.04466>
- [8] Jianzhao Huang, Xiaorui Huang, Fei Zhao, Yunpeng Liu, Hui Zhang, Fangcheng Shi, Congfeng Li, Zechen Sun, Yi Wu, Yao Hu, Yunhan Bai, and Shaosheng Cao. 2026. QP-OneModel: A Unified Generative LLM for Multi-Task Query Understanding in Xiaohongshu Search. *arXiv preprint arXiv:2602.09901* (2026). arXiv:2602.09901 <https://arxiv.org/abs/2602.09901>
- [9] Angqing Jiang, Gaoming Zhang, Jianchun Song, Kena Qi, Dayao Chen, Wei Lin, and Defu Lian. 2026. Think-to-Personalize: Unifying Reasoning and Retrieval for User-Centric Personalized Dense Retrieval. *arXiv preprint arXiv:2608.18855* (2026). arXiv:2608.18855 <https://arxiv.org/abs/2608.18855> Accepted at CIKM 2026.
- [10] Thomas Kleine Buening, Jonas Hübötter, Barna Pásztor, Idan Shenfeld, Giorgia Ramponi, and Andreas Krause. 2026. Aligning Language Models from User Interactions. In *Conference on Language Modeling*. https://self-distillation.github.io/user_interactions
- [11] Ang Li, Yufei Shi, Yuxuan Si, Yiquan Wu, Ming Cai, Xu Tan, Yi Wang, Changlong Sun, Xiaozhong Liu, and Kun Kuang. 2026. Think Then Rewrite: Reasoning Enhanced Query Rewriting for Domain Specific Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence* 40, 18 (2026), 15045–15053. doi:10.1609/aaai.v40i18.38527
- [12] Chenyi Li, Xinhui Tu, and Zaixiang Wang. 2026. DVCQR: Dual-View Conversational Query Rewriting with Stage-wise Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 22993–23014. doi:10.18653/v1/2026.acl-long.1054
- [13] Xinpeng Liu, Lu Ma, Jiayi Qiao, Mengyu Zhou, Linglong Li, Xiaofeng Bian, Haonan Chen, Xiaoxi Jiang, and Guanrun Jiang. 2026. Generative Query Suggestion via Intent Coverage and Query-Level Credit Assignment. *arXiv preprint arXiv:2609.19209* (2026). arXiv:2609.19209 <https://arxiv.org/abs/2609.19209>
- [14] Duy A. Nguyen, Rishi Kesav Mohan, Shimeng Yang, Pritom Saha Akash, and Kevin Chen-Chuan Chang. 2025. MiniELM: A Lightweight and Adaptive Query Rewriting Framework for E-Commerce Search Optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 6952–6964. doi:10.18653/v1/2025.findings-acl.363
- [15] Ruize Ou, Kai Wang, Jianzhi Shao, Tao Zhang, and Chengfu Huo. 2026. Aligning Query Rewriting with Human Cognition and Preference in E-Commerce Search. In *Proceedings of the ACM Web Conference 2026*. ACM, 7933–7942. doi:10.1145/3774904.3792806
- [16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. doi:10.52202/068431-2011
- [17] 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 Web Conference 2024*. ACM, 20–28. doi:10.1145/3589335.3648298
- [18] 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, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 53728–53741. doi:10.52202/075280-2338
- [19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017). arXiv:1707.06347 <https://arxiv.org/abs/1707.06347>
- [20] Yao Tang, Ying Yang, Ben Chen, Yufei Ma, Zihan Liang, Chenyi Lei, Wenwu Ou, and Jian Liu. 2026. OneBar: An End-to-End Content-Grounded Generative Query Recommendation Framework for E-Commerce Video Feeds. *arXiv preprint arXiv:2606.15330* (2026). arXiv:2606.15330 <https://arxiv.org/abs/2606.15330>
- [21] Mingkai Tian, Xuye, Long Meng, Liwei Chen, Zhiheng Qin, and Yi Wang. 2026. From Short Video to Clickable Search: RLVR-Enabled Listwise Query Suggestion with Retrieval-Augmented Context. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*. Association for Computational Linguistics, 552–562. doi:10.18653/v1/2026.acl-industry.38
- [22] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 9414–9423. doi:10.18653/v1/2023.emnlp-main.585
- [23] Tongyu Wen, Guanting Dong, and Zhicheng Dou. 2026. SmartSearch: Process Reward-Guided Query Refinement for Search Agents. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1971–1982. doi:10.1145/3805712.3809700
- [24] Yilin Wen, Rong Yang, Xiaojia Chang, Hong Sun, Gefu Tang, Chunhui Liu, Jeffrey Chen, Zeyu Ma, Lisong Qiu, Xiaochuan Fan, Congjia Yu, Quan Zhou, Yuheng Chen, and Zian Wang. 2026. CoRe: A Continuously Reward-Finetuned LLM Query Rewriter for Multi-Stage Context-Aware Relevance in Web-Scale Video Search. *arXiv preprint arXiv:2606.14127* (2026). arXiv:2606.14127 <https://arxiv.org/abs/2606.14127>
- [25] Zhichao Xu, Shengyao Zhuang, Xueguang Ma, Bingsen Chen, Yijun Tian, Fengran Mo, Tao Li, Jie Cao, and Vivek Srikumar. 2026. Rethinking On-policy Optimization for Query Augmentation. *Transactions on Machine Learning Research* (2026). <https://openreview.net/forum?id=mmqbjhz5Br>
- [26] Qi Zheng, Mingjie Zhong, Saisai Gong, Huimin Jiang, Kaixin Wu, Hong Liu, Jia Xu, and Linjian Mo. 2025. MAAQR: An LLM-based Multi-Agent Framework for Adaptive Query Rewriting in Alipay Search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 4289–4293. doi:10.1145/3726302.3731950
- [27] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. TTRL: Test-Time Reinforcement Learning. In *Advances in Neural Information Processing Systems*, D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (Eds.), Vol. 38. Curran Associates, Inc., 131459–131483. doi:10.52202/085713-4376