ConvSearch-R1: Enhancing Query Reformulation for Conversational Search with Reasoning via Reinforcement Learning

Changtai Zhu [1] **Siyin Wang** [1] **Ruijun Feng** [2] **Kai Song** [3] **Xipeng Qiu** [1]

Abstract

Conversational search systems require effective handling of context-dependent queries that often contain ambiguity, omission, and coreference. Conversational Query Reformulation (CQR) addresses this challenge by transforming such queries into self-contained forms suitable for standard retrieval pipelines, including those based on vector databases. Existing CQR approaches face two major limitations: a reliance on costly external supervision from human annotations or large language models, and poor alignment between the rewriting model and the downstream retriever. To address these issues, we propose ConvSearch-R1, the first self-driven framework that eliminates the need for external rewrite supervision by leveraging reinforcement learning to optimize query reformulation directly through retrieval signals obtained from vector databases. Our method introduces a novel two-stage pipeline: (1) Self-Driven Policy Warm-Up, which mitigates the cold-start problem via retrieval-guided self-distillation, and (2) Retrieval-Guided Reinforcement Learning, which employs a rank-incentive reward shaping mechanism to overcome the sparsity of traditional retrieval metrics. Extensive evaluations on the TopiOCQA and QReCC datasets show that ConvSearch-R1 significantly outperforms previous state-of-the-art methods, achieving over 10% improvement on the challenging TopiOCQA dataset while using only a 3B parameter model without any external supervision. Our results highlight the practical utility of vector databases in enabling effective, self-supervised reformulation strategies for conversational search applications.

[1]Fudan University [2]University of New South Wales [3]ByteDance Inc. Correspondence to: Xipeng Qiu <xpqiu@fudan.edu.cn>.

Proceedings of the 1st Workshop on Vector Databases at International Conference on Machine Learning, 2025.

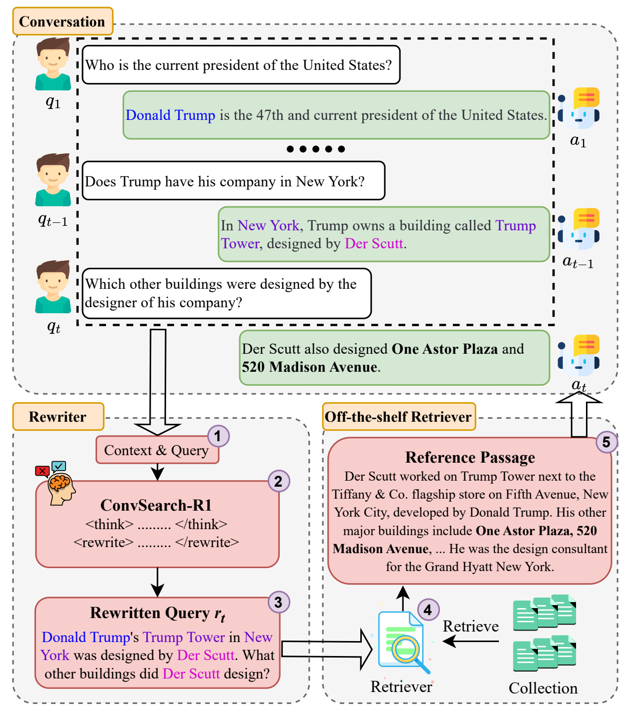


Figure 1. Illustration of the CQR task. Given a query and its context, the rewriter aims to reformulate the query into a stand-alone form, which facilitates the off-the-shelf retriever in finding the most relevant passage.

1. Introduction

Conversational search aims to fulfill users' information needs through multi-turn interactions, unlike traditional information retrieval systems that only consider single-turn, keyword-based queries (Joshi et al., 2017; Kwiatkowski et al., 2019). In these multi-turn scenarios, conversational queries often contain ambiguity, omission, and coreference, making it difficult for existing retrieval methods to accurately capture user intent (Anantha et al., 2020; Qu et al., 2020; Gao et al., 2022). Given the challenges and computational costs of training multi-turn retrievers, Conversational Query Reformulation (CQR) (Elgohary et al., 2019; Vakulenko et al., 2020; Yu et al., 2020) has emerged as a practical solution, transforming context-dependent queries into self-contained forms that can be effectively processed by off-the-shelf retrievers (As shown in Figure 1).

Existing CQR approaches have explored various strategies to address conversational search challenges. Some methods rely on explicit rewrites as supervision, obtained through either human annotations (Lin et al., 2020; Vakulenko et al., 2020; Del Tredici et al., 2021; Vakulenko et al., 2021) or knowledge distillation (Mao et al., 2023b; Mo et al., 2024a) from large language models (LLMs) like ChatGPT (OpenAI, 2023). More recent approaches attempt to leverage retrieval signals for preference optimization, though still require externally annotated data for initialization and generally limit exploration to pairwise preferences rather than optimizing for actual ranking improvements (Mo et al., 2023; Jang et al., 2024; Zhang et al., 2024; Yoon et al., 2025; Lai et al., 2025). These approaches suffer from two critical constraints: 1) the high dependency on costly and inconsistent external sources for high-quality annotation; and 2) insufficient alignment between the rewriting model and the downstream retriever. The fundamental challenge remains: how to enable query reformulation models to effectively align with retrievers without explicitly annotated reference rewrites, through self-exploration guided by retrieval feedback from vector databases.

To address these challenges, we propose ConvSearch-R1, a novel self-driven framework that completely eliminates the dependency on external rewrite supervision. Leveraging retrieval ranking signals as rewards, the model self-discovers effective rewrites through iterative exploration and exploitation. Specifically, we design a two-stage framework: (1) Self-Driven Policy Warm-Up (SDPWU), which addresses the cold-start problem by leveraging the model's few-shot reasoning capabilities combined with retrieval ranking signals from vector databases to self-distill high-quality rewrite data without external supervision; and (2) Retrieval-Guided Reinforcement Learning, which further aligns the rewrite model with the retriever through Group Relative Policy Optimization (GRPO) (Shao et al., 2024). With a carefully designed rank-incentive reward shaping, ConvSearch-R1 addresses the sparsity issue in conventional retrieval metrics (like Recall@K and NDCG@K) (Jiang et al., 2025), providing smoother learning signals rather than binary or highly skewed outcomes, enabling stable and efficient exploration of the vast reformulation space.

We validate the effectiveness and generalizability of ConvSearch-R1 through extensive experiments using 3B parameter models. Compared to approaches employing 7B models (Zhang et al., 2024; Yoon et al., 2025), our method is not only more cost-efficient but also delivers even better performance. ConvSearch-R1 achieves state-of-the-art performance on two widely-used conversational search datasets, TopiOCQA (Adlakha et al., 2021) and QReCC (Anantha et al., 2021). Notably, on the more challenging TopiOCQA dataset under dense retrieval, ConvSearch-R1, using Llama3.2-3B and Qwen2.5-3B as backbones, improves by 10.3% and 10.7% on average across all metrics, respectively, compared to previous state-of-the-art results. This demonstrates that, even without human rewrites or external distilled data, relying solely on self-distillation combined with RL under the reasoning mode enables the model to perform effectively on the CQR task.

Our contributions are summarized as follows:

- We propose ConvSearch-R1, the first conversational query rewriting approach that completely eliminates dependency on external rewrite supervision, enabling effective alignment with off-the-shelf retrievers without costly human annotations.
- We introduce a novel two-stage alignment framework comprising self-driven policy warm-up and rank-incentive reward shaping that effectively addresses the cold-start problem and reward sparsity challenges inherent in retrieval-aligned optimization.
- Our extensive experiments across multiple datasets and retrievers demonstrate substantial performance improvements over state-of-the-art methods, particularly on the challenging TopiOCQA dataset, where ConvSearch-R1 achieves over 10% average improvement across all metrics while using smaller language models and no external supervision.
- To facilitate future research in this area, we make datasets, code, and models available at `https://github.com/BeastyZ/ConvSearch-R1`.

2. Related Work

2.1. Conversational Search

Conversational search is a way of searching for information by having a natural, back-and-forth dialogue with a search system, similar to talking with a person. The core challenge of conversational search lies in addressing omission, ambiguity, and coreference present in the current query. Existing approaches to conversational search can be broadly categorized into two main types: conversational dense retrieval (CDR) and CQR. For CDR, many existing methods (Mao et al., 2024; Kim & Kim, 2022; Mo et al., 2024b) use a substantial amount of annotated session-passage pairs to fine-tune an ad-hoc retriever into a context-aware conversational retriever, a process that is costly and may not fully take the advantages of off-the-shelf retrievers. On the other hand, CQR leverages the strengths of existing ad-hoc retrieval systems by striving to transform context-dependent queries into stand-alone forms. Early studies (Voskarides et al., 2020; Lin et al., 2020) primarily relied on human rewrites to endow models with query rewriting capability. With the advent of LLMs, some recent works have begun to utilize the power of LLMs for query rewriting. Mao et al. (2023b) and Ye et al. (2023) employ ChatGPT to perform query rewriting via a training-free, purely prompt-based method. Meanwhile, many approaches (Mo et al., 2023; Zhang et al., 2024; Jang et al., 2024; Mo et al., 2024a; Lai et al., 2025; Yoon et al., 2025) distill high-quality rewrites from ChatGPT or Llama to train rewriters that possess strong rewriting ability from the outset. In contrast, ConvSearch-R1 improves query rewriting capability through self-distillation and RL-based trial and error. ConvSearch-R1 not only significantly reduces the cost of obtaining rewrites but also achieves state-of-the-art performance.

2.2. RLVR-based Retrieval

Reinforcement learning from verifiable reward (RLVR) has recently emerged as a powerful approach for enhancing language models' capabilities without explicit supervision (DeepSeek-AI et al., 2025). While recent work has integrated RLVR with retrieval mechanisms (Jin et al., 2025; Song et al., 2025; Chen et al., 2025; Sun et al., 2025), these efforts primarily focus on improving single-turn question answering by teaching LLMs to better utilize retrieval tools. In contrast, our approach leverages reasoning to optimize query reformulation specifically for conversational search contexts. In the retrieval domain, Jiang et al. (2025) applies RLVR for query generation, but differs from our work in two key aspects: (1) it uses retrieval metrics directly as rewards, while we develop a specialized rank-incentive reward function, which significantly enhances retrieval performance; and (2) it addresses only single-turn scenarios, not the conversational challenges of omission, ambiguity, and coreference resolution. Regarding conversational search specifically, existing RL approaches Wu et al. (2021); Chen et al. (2022) rely on human rewrites for reward computation, whereas our method eliminates external supervision requirements while incorporating reasoning capabilities into the rewriter.

3. ConvSearch-R1

3.1. Task Formulation

A conversational search session be defined as a sequence of alternating user queries and system answers: $S = \{(q_1, a_1), (q_2, a_2), \ldots, (q_t, a_t)\}$, where q_t denotes the user query at turn *t*, and a_t denotes the corresponding system answer. At turn *t*, the user issues a query q_t, which is potentially dependent on the previous conversational history $C_{t-1} = \{(q_1, a_1), (q_2, a_2), \ldots, (q_{t-1}, a_{t-1})\}$. The task of CQR is, given the current query q_t and the preceding conversational history C_{t-1}, to generate a context-independent reformulated query $r_t = f(q_t, C_{t-1})$, where $f(\cdot)$ is the reformulation function that leverages both previous queries and system answers to resolve omission, ambiguity, and coreference in q_t. The optimal reformulation function f^* can be defined as:

$$f^* = \arg\max_f \mathbb{E}_S[\delta(\mathcal{T}(r_t), \hat{p}_t)],$$

where $\hat{p}_t$ is the gold-standard passage for q_t in the collection, $\mathcal{T}(\cdot)$ is the retriever that returns passages given a query, and $\delta(\cdot, \cdot)$ is a matching function that measures the relevance between the retrieved passage and the gold passage.

3.2. Overview

ConvSearch-R1 employs a reasoning-based approach to enable the rewriter to fully grasp the omission, ambiguity, and coreference present in the current user query, thereby generating rewrites that are both context-independent and semantically rich. As illustrated in Figure 2, ConvSearch-R1 does not require any external supervised data (reference rewrite) and consists of the following stages: (1) Self-distill a set of format-compliant data from the π_{init} through few-shot learning, aligning it with the preferences of the retriever. Only the top-ranked (rank-1) rewrites are retained and used to fine-tune the π_{init}, thus getting the π_{SFT} with initial ability to follow the desired output format and perform query rewriting; (2) Further improve the π_{SFT} using RL with rank-incentive reward shaping, yielding the final rewriter π_θ. The output format of the rewriter is required to strictly adhere to the following structure: *<think>reasoning process here </think>\n<rewrite>rewrite here </rewrite>*.

3.3. Self-Driven Policy Warm-Up

Following Chu et al. (2025)'s findings on SFT's crucial role in stabilizing output formats for effective RL training, we introduce a self-driven policy warm-up strategy that avoids costly knowledge distillation from more powerful LLMs used in previous approaches.

Self-Distillation. We begin by generating rewrites using few-shot prompting applied to our initial model π_{init}:

$$D^d = \{y_i = \pi_{init}(x_i, \rho, instruction) \mid x_i \in D\},$$

where D is the original dataset, ρ is a fixed set of few-shot examples, and D^d is the distilled dataset. We then filter these outputs to obtain format-compliant samples: $D^f = \{y \in D^d \mid g(y) = 1\}$, where $g : Y \to \{0, 1\}$ is a format validation function indicating whether the output meets the required constraints.

Preference Collection and Supervised Fine-Tuning. We identify high-quality rewrites by retaining only samples

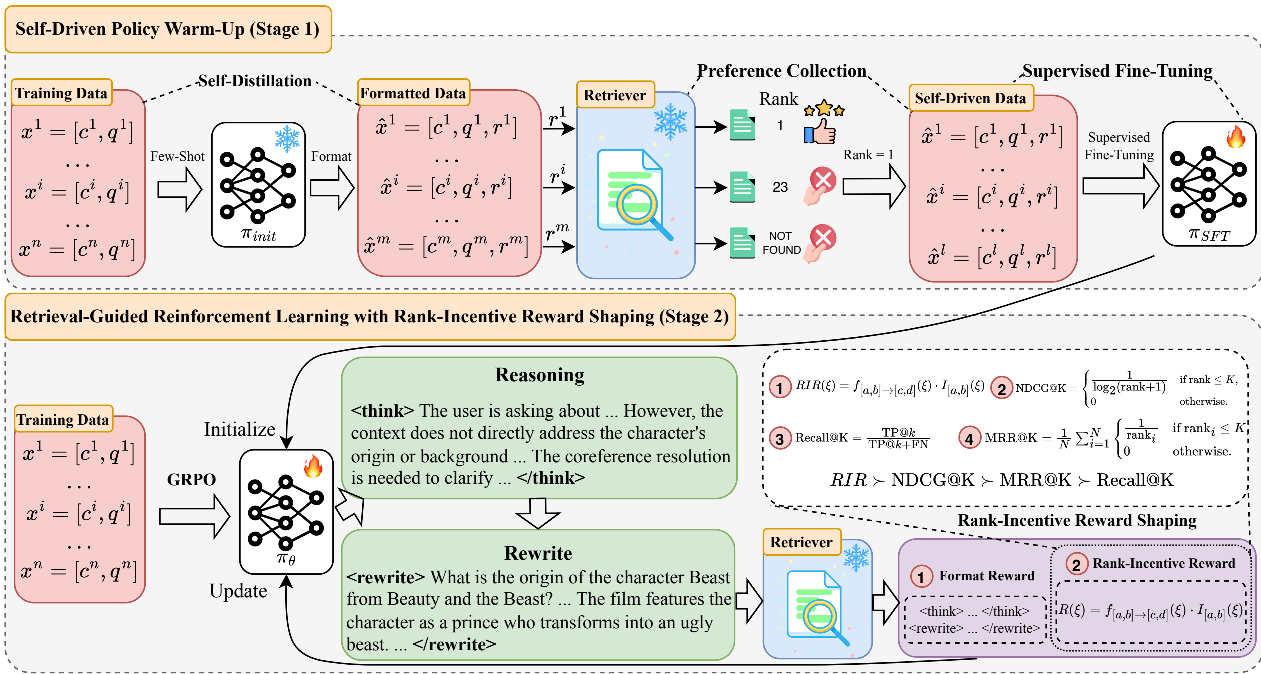


Figure 2. Overview of ConvSearch-R1. In Stage 1, we self-distilled a set of high-quality data using few-shot learning and obtained the corresponding SFT model. In Stage 2, we further improved the rewriter's performance via RL by refining the reward function.

whose rewrites rank the gold passage at position 1, creating our self-driven data (SD-DATA). Each sample in SD-DATA is represented as a triplet $[c,q,r]$, containing conversational history c, current query q, and the reasoning-rewrite pair r. For more details about data collection, see Appendix B. Finally, we fine-tune π_{init} on SD-DATA to maximize the likelihood of ground-truth outputs, producing π_{SFT} with fundamental capabilities in format adherence, reasoning, and query rewriting. Through this process, our model learns to first generate an appropriate reasoning process and then produce a context-independent rewrite, conditioned on both the conversation history and the generated reasoning.

3.4. Retrieval-Guided Reinforcement Learning

Rank-Incentive Reward Shaping. The design of reward functions is critical in RL, directly impacting the learning effectiveness of policy models (Devidze et al., 2021). Unlike Jiang et al. (2025) which directly uses retrieval metrics as reward signals, we propose Rank-Incentive Reward Shaping (RIRS) to address the reward sparsity problem. As shown in Appendix Figure 8, directly using metrics like MRR@3 and NDCG@3 as rewards leads to severe reward sparsity, hindering effective model learning.

RIRS utilizes retrieval ranking positions to create a more informative reward signal rather than relying on binary retrieval metrics. Considering users typically pay more attention to top positions, RIRS implements a piecewise reward function that allocates differentiated reward intensities—assigning higher rewards for top positions (1-10) while maintaining proportionally smaller but meaningful rewards for moderate positions (11-100). This approach preserves semantic thresholds of retrieval quality while ensuring dense feedback signals throughout the policy optimization process. The RIRS reward function is formally defined as:

$$\begin{aligned} R(\xi) = &f_{[1,10]\to[1,2]}(\xi) \cdot I_{[1,10]}(\xi) \\ &+ f_{[10,100]\to[0,1]}(\xi) \cdot I_{(10,100]}(\xi), \end{aligned} \tag{1}$$

where $f_{A\to B}$ represents a function mapping from interval A to interval B, $I_A(\xi)$ is the indicator function, which equals 1 when ξ is in set A and 0 otherwise, ξ is the rank variable.

Considering format correctness, the complete reward function is:

$$R(\xi,\phi) = R(\xi) \cdot I(\phi = 1) + \delta \cdot I(\phi = 0), \tag{2}$$

where $\phi \in 0, 1$ is the format compliance indicator, and $\delta = -0.1$ is the penalty term for format non-compliance.

GRPO. Equipped with the rewriter π_{SFT} obtained from Stage 1 SDPWU and the carefully designed reward function RIR, we adopt GRPO (Shao et al., 2024) as the specific RL algorithm. GRPO is an efficient algorithm that eliminates the need for an explicit reward model and value model.

Through RL with Rank-Incentive Reward Shaping, an existing π_θ is incentivized to explore the solution space while rollouts $\{r_i\}_{i=1} \sim \pi_{\theta_{old}}(\cdot|x)$ are used to maximize retrieval performance. The optimization objective is formulated as:

$$J_{GRPO}(\theta) = \mathbb{E}\Big[\sum_i \min\Big(\frac{\pi_\theta(r_i|x)}{\pi_{\theta_{old}}(r_i|x)}A_i,\ \text{clip}\Big(\frac{\pi_\theta(r_i|x)}{\pi_{\theta_{old}}(r_i|x)}, 1-\epsilon, 1+\epsilon\Big)A_i\Big) - \beta D_{KL}(\pi_\theta||\pi_{SFT})\Big], \tag{3}$$

where $A_i = \frac{R(\xi_i,\phi_i)-\text{mean}(R(\xi,\phi))}{\text{std}(R(\xi,\phi))}$ represents the normalized advantage of the i-th rewritten query within the current group, calculated using our RIRS reward function. The parameter ϵ controls the clipping threshold, while β regulates the KL divergence penalty.

4. Experimental Setup

Settings. We follow prior work (Yoon et al., 2025) in configuring the datasets, retrievers, and metrics. For datasets, we employ TopiOCQA (Adlakha et al., 2021) and QReCC (Anantha et al., 2021), which are widely used in conversational search task. For retrievers, we utilize BM25 as the sparse retriever for all experiments and ANCE (Xiong et al., 2020) as the dense retriever for all experiments, where ANCE is trained on the MS-MARCO (Campos et al., 2016) passage retrieval tasks. Notably, we do not train any retriever in our experiments. For metrics, we adopt MRR@3, NDCG@3, and Recall@K (referred to as R@K in this paper) for evaluation. See Appendix E.1 & E.2 for more details.

Baselines[1]**.** We consider four categories of baselines in our experiments. The first category comprises basic baselines that do not involve any query rewriting optimization, including Human Rewrite, Raw, DS-R1-Distill-Qwen-7B, Llama3.2-3B, and Qwen2.5-3B. The second category consists of baselines that fine-tune small-scale models (e.g., T5 and BERT), including QuReTec (Voskarides et al., 2020), T5QR (Lin et al., 2020), CONQRR (Wu et al., 2021), ConvGQR (Mo et al., 2023), EDIRCS (Mao et al., 2023a), IterCQR (Jang et al., 2024), ADACQR (Lai et al., 2025), and CHIQ-Fusion (Mo et al., 2024a). The third category involves baselines that fine-tune LLMs (e.g., Llama2-7B) , including RETPO (Yoon et al., 2025) and AdaQR (Zhang et al., 2024). The fourth category is training-free, leveraging prompt-based methods with ChatGPT for query rewriting. This category includes LLM4CS (Mao et al., 2023b) and InfoCQR (Ye et al., 2023). A detailed description of each aforementioned baseline is presented in Appendix E.3.

Implementation Details. We employ Llama3.2-3B and Qwen2.5-3B as the backbone models for our rewrite. For training, we utilize verl (Sheng et al., 2024), a flexible and efficient RLHF framework. The BM25 retriever is implemented using Pyserini (Lin et al., 2021), while the ANCE retriever is built with Faiss (Johnson et al., 2017). Evaluation metrics are computed with pytrec_eval (Gysel & de Rijke, 2018). More implementation details can be found in Appendix E.4.

[1]We focus exclusively on the CQR and do not consider methods designed for CDR.

5. Results and Analysis

5.1. Main Results

Table 1 shows the retrieval performance of various methods in dense and sparse settings, leading to the following conclusions:

Employing a two-stage alignment framework further enhances the final performance in the absence of external supervised data. Directly using the current user query (`Raw`) for retrieval yields the poorest performance, underscoring the necessity of CQR. Notably, `Human Rewrite` not only fails to deliver optimal results but also performs significantly worse than many other baselines. This suggests that, in addition to the high annotation costs, human rewrites are not aligned with retriever preferences. Furthermore, directly applying the reasoning model (i.e., `DS-R1-Distill-Qwen-7B`) to conversational search yields even worse results than `Human Rewrite`. This indicates that straightforward general-domain reasoning is not a silver bullet, and even models equipped with Long-CoT capability (i.e., R1-like models) struggle to excel in conversational search scenarios. Methods that rely solely on human rewrites as supervision signals (e.g., `QuReTeC`, `T5QR`) exhibit the lowest performance among all baselines, highlighting the clear limitations of human rewrites. In contrast, approaches that utilize retrieval signals for supervision (e.g., `CONQRR`, `EDIRCS`, `IterCQR`, `ADACQR`, `RETPO`, `AdaQR`) achieve notable performance improvements, yet still fall short of our method. These results collectively demonstrate the effectiveness of our proposed two-stage alignment framework. Notably, compare to the baselines, ConvSearch-R1 achieves new state-of-the-art performance across most experimental settings without the need for any external supervised data.

While the baselines leverage larger language models to achieve competitive performance, our 3B rewriter demonstrates superior results. Many studies (e.g., `RETPO`, `AdaQR`, `LLM4CS`, `InfoCQR`) have attempted to tackle the CQR task by leveraging larger language models (such as ChatGPT). However, our 3B-parameter model significantly outperforms these approaches and achieves new state-of-the-art results. Both `RETPO` and `AdaQR` employ Direct Preference Optimization (Rafailov et al., 2023) to align the retriever's preferences, but this strategy overlooks the potential of enabling the model to fit preferences through

Table 1. Results of both dense and sparse retrieval on TopiOCQA and QReCC. **NE** denotes no external distillation, indicating that no external data was distilled from open-source or closed-source LLMs. **NH** stands for no human, meaning that no human rewrites were utilized. **Bold** indicates the best result, and the rest of the tables follow the same convention. Grey indicates the improvements over SOTA baselines.

Type	Method	NE	NH	TopiOCQA				QReCC				Avg
				MRR@3	NDCG@3	R@10	R@100	MRR@3	NDCG@3	R@10	R@100	
Sparse (BM25)	Human Rewrite	-	-	-	-	-	-	39.8	36.3	62.7	**98.5**	-
	Raw	-	-	2.1	1.8	4.0	9.2	6.5	5.5	11.1	21.5	7.7
	DS-R1-Distill-Qwen-7B	-	-	10.0	8.6	18.8	38.3	29.4	27.3	44.0	63.9	30.0
	QuReTeC *(SIGIR 2020)*	✓	✗	8.5	7.3	16.0	-	34.0	30.5	55.5	-	-
	T5QR	✓	✗	11.3	9.8	22.1	44.7	33.4	30.2	53.8	86.1	36.4
	CONQRR *(EMNLP 2022)*	✗	✗	-	-	-	-	38.3	-	60.1	88.9	-
	ConvGQR *(ACL 2023)*	✓	✗	12.4	10.7	23.8	45.6	44.1	41.0	64.4	88.0	41.3
	EDIRCS *(ACL 2023)*	✗	✗	-	-	-	-	41.2	-	62.7	90.2	-
	IterCQR *(NAACL 2024)*	✗	✓	16.5	14.9	29.3	54.1	46.7	44.1	64.4	85.5	44.4
	ADACQR *(COLING 2025)*	✗	✓	28.3	26.5	48.9	71.2	55.1	52.5	76.5	93.7	56.6
	CHIQ-Fusion *(EMNLP 2024)*	✗	✓	25.6	23.5	44.7	-	54.3	51.9	**78.5**	-	-
	RETPO *(NAACL 2025)*	✗	✓	28.3	26.5	48.3	73.1	50.0	47.3	69.5	89.5	54.1
	AdaQR *(EMNLP 2024)*	✗	✓	20.3	18.0	37.1	66.2	50.6	48.0	69.6	-	-
	LLM4CS *(EMNLP 2023)*	✗	✓	18.9	17.7	33.7	-	47.8	45,0	69.1	-	-
	InfoCQR *(EMNLP 2023)*	✗	✓	-	-	-	-	48.9	46.3	66.4	-	-
	LLama3.2-3B	-	-	4.8	4.0	8.6	20.7	22.5	21.0	33.7	49.8	20.6
	+ ConvSearch-R1(ours)	✓	✓	**37.8**	**36.2**	**59.6**	**80.1**	**55.9**	**54.3**	77.2	89.0	**61.3**
				+9.5	*+9.7*	*+10.7*	*+7.0*	*+0.8*	*+1.8*	*-1.3*	*-4.7*	*+4.7*
	Qwen2.5-3B	-	-	8.8	7.5	17.3	36.1	27.3	25.0	42.3	64.6	28.6
	+ ConvSearch-R1(ours)	✓	✓	**35.2**	**33.5**	**57.8**	**79.9**	**56.5**	**54.8**	76.3	88.1	**60.3**
				+6.9	*+7.0*	*+8.9*	*+6.8*	*+1.4*	*+2.3*	*-2.2*	*-5.6*	*+3.7*
Dense (ANCE)	Human Rewrite	-	-	-	-	-	-	38.4	35.6	58.6	78.1	-
	Raw	-	-	4.1	3.8	7.5	13.8	10.2	9.3	15.7	22.7	10.9
	DS-R1-Distill-Qwen-7B	-	-	21.0	19.9	36.4	53.4	28.4	25.9	43.9	59.0	36.0
	QuReTeC *(SIGIR 2020)*	✓	✗	11.2	10.5	20.2	-	35.0	32.6	55.0	-	-
	T5QR	✓	✗	23.0	22.2	37.6	54.4	34.5	31.8	53.1	72.8	41.2
	CONQRR *(EMNLP 2022)*	✗	✗	-	-	-	-	41.8	-	65.1	84.7	-
	ConvGQR *(ACL 2023)*	✓	✗	25.6	24.3	41.8	58.8	42.0	39.1	63.5	81.8	47.1
	EDIRCS *(ACL 2023)*	✗	✗	-	-	-	-	42.1	-	65.6	**85.3**	-
	IterCQR *(NAACL 2024)*	✗	✓	26.3	25.1	42.6	62.0	42.9	40.2	65.5	84.1	48.6
	ADACQR *(COLING 2025)*	✗	✓	38.5	37.6	58.4	75.0	45.8	42.9	67.3	83.8	56.2
	CHIQ-Fusion *(EMNLP 2024)*	✗	✓	38.0	37.0	61.6	-	47.2	44.2	**70.7**	-	-
	RETPO *(NAACL 2025)*	✗	✓	30.0	28.9	49.6	68.7	44	41.1	66.7	84.6	51.7
	AdaQR *(EMNLP 2024)*	✗	✓	38.1	36.6	61.3	79.9	43.4	40.8	65.6	-	-
	LLM4CS *(EMNLP 2023)*	✗	✓	27.7	26.7	43.3	-	44.8	42.1	66.4	-	-
	InfoCQR *(EMNLP 2023)*	✗	✓	-	-	-	-	43.9	41.3	65.6	-	-
	LLama3.2-3B	-	-	12.8	11.7	22.6	38.1	19.1	17.4	30.4	44.1	24.5
	+ ConvSearch-R1(ours)	✓	✓	**50.5**	**50.1**	**72.0**	**86.3**	**50.2**	**48.1**	70.6	82.8	**63.8**
				+12.0	*+12.5*	*+10.4*	*+6.4*	*+3.0*	*+3.9*	*-0.1*	*-2.5*	*+7.6*
	Qwen2.5-3B	-	-	19.2	18.3	33.0	46.7	29.3	27.2	44.2	59.0	34.6
	+ ConvSearch-R1(ours)	✓	✓	**51.4**	**51.3**	**72.0**	**85.7**	**49.7**	**47.7**	69.8	81.6	**63.7**
				+12.9	*+13.7*	*+10.4*	*+5.8*	*+2.5*	*+3.5*	*-0.9*	*-3.7*	*+7.5*

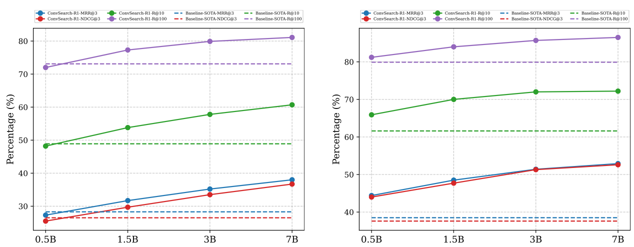
trial and error. In contrast, ConvSearch-R1 utilizes RL with a Rank-Incentive Reward, which allows the model to explore and refine preference alignment through iterative trial and error, ultimately leading to a more optimal solution for retriever preference modeling with reduced parameter usage.

5.2. Generalization on Unseen Datasets

For the evaluation of generalization ability, we trained models on the TopiOCQA training set and evaluated them on the QReCC test set, and vice versa. As shown in Table 2, our method demonstrates superior generalization performance

Table 2. Performance on unseen datasets.

Method	TopiOCQA		QReCC	
	MRR@3	NDCG@3	MRR@3	NDCG@3
Sparse(BM25)				
IterCQR	13.7	12.2	44.9	42.4
ADACQR	14.0	12.6	-	-
RETPO	17.2	-	40.1	-
ConvSearch-R1 (LLama)	**27.6**	**25.8**	**49.9**	**47.6**
ConvSearch-R1 (Qwen)	**24.0**	**22.1**	**46.2**	**43.9**
Dense(ANCE)				
IterCQR	17.8	16.4	40.1	37.4
ADACQR	20.1	18.6	-	-
RETPO	23.2	-	40.9	-
ConvSearch-R1 (LLama)	**36.8**	**35.4**	**42.8**	**40.2**
ConvSearch-R1 (Qwen)	**35.2**	**34.2**	**41.5**	**39.0**



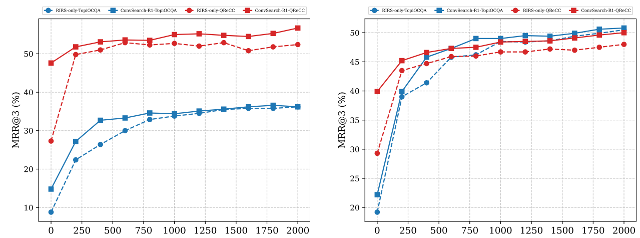
(a) Comparison on BM25 (b) Comparison on ANCE

Figure 3. Model scale analysis on TopiOCQA using Qwen2.5 Series.

compared to other approaches. This significant improvement in generalization can be primarily attributed to our use of RL with a Rank-Incentive Reward. By employing a well-designed reward function, our training paradigm encourages the model to interact with a broader and more diverse set of high-quality data through trial and error, thereby enhancing its ability to generalize across different datasets. See Appendix Table 8 for a full comparison.

5.3. Scaling Behavior of Model Performance

To further validate the generalizability of our approach across models of varying parameter scales, we conducted experiments on the TopiOCQA dataset using the Qwen2.5 Series models, as shown in Figure 3. Our results demonstrate that our method consistently outperforms SOTA baseline across almost all model sizes. Notably, as the model size increases, the performance gap between our method and the SOTA baseline becomes increasingly pronounced. Interestingly, even when using a relatively small 0.5B parameter model under dense (ANCE) retrieval, our approach still significantly surpasses the SOTA baseline. These findings indicate that ConvSearch-R1 exhibits strong generalizability across models of any scale.



(a) Comparison on BM25 (b) Comparison on ANCE

Figure 4. Warm-up analysis using Qwen2.5-3B

5.4. Ablation Study

Overall Analysis. ConvSearch-R1 is a two-stage alignment framework. In Stage 1 (§ 3.3), an initial preference-aligned rewriter is obtained through self-distillation and preference-based filtering. In Stage 2 (§ 3.4), reinforcement learning with carefully designed rewards is employed to further align the rewriter with retriever preferences. To validate the effectiveness of each component, we conduct ablation studies. As shown in Table 3, each component in ConvSearch-R1 plays a critical role across different retrievers and datasets, demonstrating the rationality and effectiveness of our framework's design.

Warm-Up Analysis. To validate the necessity of Stage 1 (§ 3.3), we selected checkpoints from various steps during the training process using different retrievers and on different datasets. These checkpoints were then evaluated accordingly. As shown in Figure 4, as the number of training steps increases, the MRR@3 score gradually improves and eventually reaches a plateau. At every training step, ConvSearch-R1 consistently outperforms the RIRS-only model (i.e., the model trained without Stage 1) in terms of MRR@3. These results provide strong evidence for the essential role of Stage 1 in enhancing model performance.

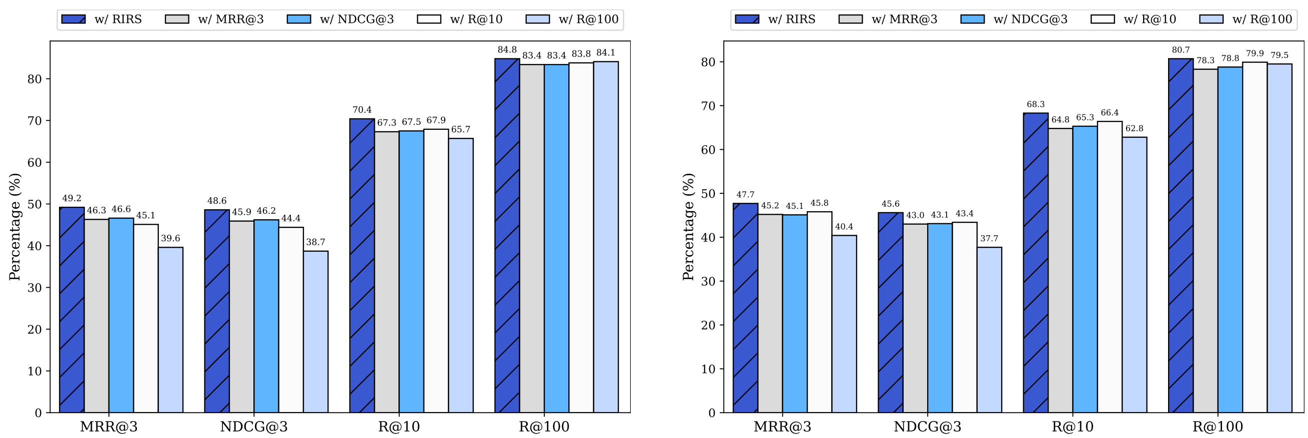
Reward Analysis. To further validate the necessity of reward shaping in stage 2 (§ 3.4), we conducted comparative experiments on different reward functions using the Qwen2.5-3B model with a dense retriever architecture. As shown in Figure 5, the rewriter utilizing Rank-Incentive Reward Shaping consistently achieved the best performance across all settings. These results provide strong evidence for the necessity of reward shaping in our task.

5.5. Case Study

Why does ConvSearch-R1 lead to significant performance improvements? To answer this question, we analyze a specific case selected from the QReCC test set, as shown in Appendix Table 9. In this case, the user asks two questions within a single query, both involving coref-

Table 3. Ablation results for both dense and sparse retrieval on the TopiOCQA and QReCC with full metrics. **SDPWU** and **RIRS** denote the stage1 (§ 3.3) and stage2 (§ 3.4) of ConvSearch-R1, respectively.

Type	Method	SDPWU	RIRS	TopiOCQA				QReCC			
				MRR@3	NDCG@3	R@10	R@100	MRR@3	NDCG@3	R@10	R@100
Sparse(BM25)	Qwen2.5-3B	✗	✗	8.8	7.5	17.3	36.1	27.3	25.0	42.3	64.6
	+ w/ SDPWU	✓	✗	14.8	13.0	27.8	50.9	47.6	45.5	64.3	79.8
	+ w/ RIRS	✗	✓	32.6	31.0	56.2	79.0	51.7	49.9	71.9	85.8
	+ ConvSearch-R1	✓	✓	**35.2**	**33.5**	**57.8**	**79.9**	**56.5**	**54.8**	**76.3**	**88.1**
Dense(ANCE)	Qwen2.5-3B	✗	✗	19.2	18.3	33.0	46.7	29.3	27.2	44.2	59.0
	+ w/ SDPWU	✓	✗	22.2	20.6	38.7	58.3	39.9	37.5	59.3	73.8
	+ w/ RIRS	✗	✓	49.2	48.6	70.4	84.8	47.7	45.6	68.3	80.7
	+ ConvSearch-R1	✓	✓	**51.4**	**51.3**	**72.0**	**85.7**	**49.7**	**47.7**	**69.8**	**81.6**



(a) Comparison on TopiOCQA

(b) Comparison on QReCC

Figure 5. Comparison of different reward functions for dense retrieval on the TopiOCQA and QReCC datasets.

erence to previous context. In the early stages of training, the alignment between the rewriter and the retriever's preferences is limited. We observe that, during the reasoning process, the rewriter abandons the reformulation of the first question, resulting in a final rewrite that only addresses the second question. Consequently, the gold passage is not retrieved within the top 100 results. In the later stages of training, after extensive trial and error, the rewriter achieves a much higher degree of alignment with the retriever's preferences. We find that the rewriter repeatedly considers both user questions during reasoning and generates a rewrite that successfully resolves the coreference for both. Notably, the rewriter even generates pseudo-passages to supplement missing information. Thanks to this comprehensive consideration and the generation of pseudo-passages, the rewriter is able to retrieve the gold passage at the top-1 position for this case in the later training stage. These findings provide strong evidence that our method can effectively align with the retriever's preferences and achieve state-of-the-art performance on the benchmark.

6. Conclusion

In this paper, we propose ConvSearch-R1, a novel two-stage alignment framework for CQR that operates without the need for any external supervised data (reference rewrite). Specifically, we employ a self-distillation and rank-based data filtering strategy to construct high-quality training data for SFT. Subsequently, we leverage RL with rank-incentive reward shaping to further refine the model through trial and error. This enables the rewriter to align its output with the retriever's preferences through reasoning process. To the best of our knowledge, we are the first to tackle the CQR task without relying on any form of external supervision. Experimental results demonstrate that ConvSearch-R1 achieves state-of-the-art performance on both in-domain and out-of-domain datasets. Furthermore, our method exhibits a consistent scaling law with respect to model size, and ablation studies confirm that each component of our framework contributes substantially to the overall performance improvement.

References

Adlakha, V., Dhuliawala, S., Suleman, K., de Vries, H., and Reddy, S. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483, 2021. URL https://api.semanticscholar.org/CorpusID:238259221.

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S. G., and Chappidi, S. Open-domain question answering goes conversational via question rewriting. In *North American Chapter of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:222290679.

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., and Chappidi, S. Open-domain question answering goes conversational via question rewriting. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 520–534, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.44. URL https://aclanthology.org/2021.naacl-main.44/.

Campos, D. F., Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L., and Mitra, B. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268, 2016. URL https://api.semanticscholar.org/CorpusID:1289517.

Chen, M., Li, T., Sun, H., Zhou, Y., Zhu, C., Wang, H., Pan, J. Z., Zhang, W., zeng Chen, H., Yang, F., Zhou, Z., and Chen, W. Research: Learning to reason with search for llms via reinforcement learning. *ArXiv*, abs/2503.19470, 2025. URL https://api.semanticscholar.org/CorpusID:277313597.

Chen, Z., Zhao, J., Fang, A., Fetahu, B., Rokhlenko, O., and Malmasi, S. Reinforced question rewriting for conversational question answering. In Li, Y. and Lazaridou, A. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 357–370, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-industry.36. URL https://aclanthology.org/2022.emnlp-industry.36/.

Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., and Ma, Y. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *ArXiv*, abs/2501.17161, 2025. URL https://api.semanticscholar.org/CorpusID:275932560.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J.-M., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B.-L., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D.-L., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S.-K., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W.-X., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X.-C., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y.-J., He, Y., Xiong, Y., Luo, Y.-W., mei You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., guo Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z.-A., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948, 2025. URL https://api.semanticscholar.org/CorpusID:275789950.

Del Tredici, M., Barlacchi, G., Shen, X., Cheng, W., and de Gispert, A. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 2974–2978, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482164. URL https://doi.org/10.1145/3459637.3482164.

Devidze, R., Radanovic, G., Kamalaruban, P., and Singla, A. Explicable reward design for reinforcement learning agents. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Elgohary, A., Peskov, D., and Boyd-Graber, J. Can you unpack that? learning to rewrite questions-in-context. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5918–5924, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1605. URL https://aclanthology.org/D19-1605/.

Gao, J., Xiong, C., Bennett, P. N., and Craswell, N. Neural approaches to conversational information retrieval. *Neural Approaches to Conversational Information Retrieval*, 2022. URL https://api.semanticscholar.org/CorpusID:245986359.

Gysel, C. V. and de Rijke, M. Pytrec_eval: An extremely fast python interface to trec_eval. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018. URL https://api.semanticscholar.org/CorpusID:13691943.

Jang, Y., Lee, K.-i., Bae, H., Lee, H., and Jung, K. IterCQR: Iterative conversational query reformulation with retrieval guidance. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8121–8138, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.449. URL https://aclanthology.org/2024.naacl-long.449/.

Jiang, P., Lin, J., Cao, L., Tian, R., Kang, S., Wang, Z., Sun, J., and Han, J. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *ArXiv*, abs/2503.00223, 2025. URL https://api.semanticscholar.org/CorpusID:276742133.

Jin, B., Zeng, H., Yue, Z., Wang, D., Zamani, H., and Han, J. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *ArXiv*, abs/2503.09516, 2025. URL https://api.semanticscholar.org/CorpusID:276937772.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2017. URL https://api.semanticscholar.org/CorpusID:926364.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551, 2017. URL https://api.semanticscholar.org/CorpusID:26501419.

Kim, S. and Kim, G. Saving dense retriever from shortcut dependency in conversational search. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10278–10287, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.701. URL https://aclanthology.org/2022.emnlp-main.701/.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Lai, Y., Wu, J., Zhang, C., Sun, H., and Zhou, D. AdaCQR: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7698–7720, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.515/.

Lin, J. J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *ArXiv*, abs/2102.10073, 2021. URL https://api.semanticscholar.org/CorpusID:231979055.

Lin, S.-C., Yang, J.-H., Nogueira, R., Tsai, M.-F., Wang, C.-J., and Lin, J. J. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *ArXiv*, abs/2004.01909, 2020. URL https://api.semanticscholar.org/CorpusID:214802570.

Mao, K., Dou, Z., Liu, B., Qian, H., Mo, F., Wu, X., Cheng, X., and Cao, Z. Search-oriented conversational query editing. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4160–4172, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.256. URL https://aclanthology.org/2023.findings-acl.256/.

Mao, K., Dou, Z., Mo, F., Hou, J., Chen, H., and Qian, H. Large language models know your contextual search intent: A prompting framework for conversational search. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1211–1225, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.86. URL https://aclanthology.org/2023.findings-emnlp.86/.

Mao, K., Deng, C., Chen, H., Mo, F., Liu, Z., Sakai, T., and Dou, Z. ChatRetriever: Adapting large language models for generalized and robust conversational dense retrieval. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1227–1240, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.71. URL https://aclanthology.org/2024.emnlp-main.71/.

Mo, F., Mao, K., Zhu, Y., Wu, Y., Huang, K., and Nie, J.-Y. ConvGQR: Generative query reformulation for conversational search. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4998–5012, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.274. URL https://aclanthology.org/2023.acl-long.274/.

Mo, F., Ghaddar, A., Mao, K., Rezagholizadeh, M., Chen, B., Liu, Q., and Nie, J.-Y. CHIQ: Contextual history enhancement for improving query rewriting in conversational search. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2268, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.135. URL https://aclanthology.org/2024.emnlp-main.135/.

Mo, F., Qu, C., Mao, K., Zhu, T., Su, Z., Huang, K., and Nie, J.-Y. History-aware conversational dense retrieval. In *Annual Meeting of the Association for Computational Linguistics*, 2024b. URL https://api.semanticscholar.org/CorpusID:267321005.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W. B., and Iyyer, M. Open-retrieval conversational question answering. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. URL https://api.semanticscholar.org/CorpusID:218869571.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023. URL https://api.semanticscholar.org/CorpusID:258959321.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J.-M., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. *Proceedings of the Twentieth European Conference on Computer Systems*, 2024. URL https://api.semanticscholar.org/CorpusID:272987758.

Song, H., Jiang, J., Min, Y., Chen, J., Chen, Z., Zhao, W. X., Fang, L., and Wen, J.-R. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *ArXiv*, abs/2503.05592, 2025. URL https://api.semanticscholar.org/CorpusID:276884818.

Sun, H., Qiao, Z., Guo, J., Fan, X., Hou, Y., Jiang, Y., Xie, P., Huang, F., and Zhang, Y. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*, 2025.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. Question rewriting for conversational question answering. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2020. URL https://api.semanticscholar.org/CorpusID:216868342.

Vakulenko, S., Voskarides, N., Tu, Z., and Longpre, S. A comparison of question rewriting methods for conversational passage retrieval. *CoRR*, abs/2101.07382, 2021. URL https://arxiv.org/abs/2101.07382.

Voskarides, N., Li, D., Ren, P., Kanoulas, E., and de Rijke, M. Query resolution for conversational search with limited supervision. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. URL https://api.semanticscholar.org/CorpusID:218840824.

Wu, Z., Luan, Y., Rashkin, H., Reitter, D., and Tomar, G. S. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:245218563.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *ArXiv*, abs/2007.00808, 2020. URL https://api.semanticscholar.org/CorpusID:220302524.

Ye, F., Fang, M., Li, S., and Yilmaz, E. Enhancing conversational search: Large language model-aided informative query rewriting. *ArXiv*, abs/2310.09716, 2023. URL https://api.semanticscholar.org/CorpusID:264145935.

Yoon, C., Kim, G., Jeon, B., Kim, S., Jo, Y., and Kang, J. Ask optimal questions: Aligning large language models with retriever‘s preference in conversation. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5899–5921, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.328/.

Yu, S. Y., Liu, J., Yang, J., Xiong, C., Bennett, P. N., Gao, J., and Liu, Z. Few-shot generative conversational query rewriting. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. URL https://api.semanticscholar.org/CorpusID:219559295.

Zhang, T., Li, K., Luo, H., Wu, X., Glass, J. R., and Meng, H. M. Adaptive query rewriting: Aligning rewriters through marginal probability of conversational answers. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13444–13461, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.746. URL https://aclanthology.org/2024.emnlp-main.746/.

A. ConvSearch-R1 VS Prior Works

As depicted in Figure 9, the distinction of ConvSearch-R1 from previous studies lies in the approach to obtaining rewrites. Prior research has always relied on external sources (e.g., human rewrites or powerful LLMs) to generate rewrites and has failed to endow the rewriter with reasoning capability. In contrast, ConvSearch-R1 acquires rewrites through self-distillation and RL, utilizing trial and error, and equips the rewriter with reasoning ability.

B. Data Collection

We performed self-distillation on Qwen2.5-3B and Llama3.2-3B using the prompts specified in Figure 6 on the training set of TopiOCQA (Adlakha et al., 2021) and QReCC (Anantha et al., 2021). We retained only those samples that conformed to the required format and for which the rewritten query resulted in the gold passage being ranked first. The number of qualified samples for each model on the TopiOCQA and QReCC datasets is summarized in Table 4.

Table 4. The number of samples obtained using the SDPWU (§ 3.3) on TopiOCQA and QReCC datasets for both Qwen2.5-3B and Llama3.2-3B models.

Model	TopiOCQA	QReCC
	Sample Nums	Sample Nums
Qwen2.5-3B	5892	7759
Llama3.2-3B	6034	6623

Prompt for few-shot learning:

Given a query and its context, you must first think about the reasoning process in the mind to decontextualize the query by resolving coreference and omission issues. Then, provide the user with a rewrite that retains its original meaning and is as informative as possible to help search engines retrieve relevant documents effectively. The reasoning process and rewrite should be enclosed within <think></think>and <rewrite></rewrite>tags, respectively, i.e., <think>reasoning process here </think>\n<rewrite>rewrite here </rewrite>.

Here is an example for your reference:
Example Begin
{*example*}
Example End

Context Begin
{*context*}
Context End

Query: {*query*}

Figure 6. Prompt for few-shot learning

C. Results on Llama2-7B

To ensure a fair comparison, we followed the experimental protocols of Yoon et al. (2025) (using Llama2-7B) and Zhang et al. (2024) (using Mistral-7B), applying our method on the Llama2-7B[2] model under the same experimental settings. As shown in Table 5, ConvSearch-R1 consistently achieves state-of-the-art performance on Llama2-7B. This result provides strong evidence for the effectiveness and generalizability of ConvSearch-R1 across different models. It is worth noting that Llama2-7B heavily relies on SFT to acquire robust format adherence capability; without SFT, the stability of RL training can be significantly compromised, resulting in worse performance.

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Prompt for training and inference:

Given a query and its context, you must first think about the reasoning process in the mind to decontextualize the query by resolving coreference and omission issues. Then, provide the user with a rewrite that retains its original meaning and is as informative as possible to help search engines retrieve relevant documents effectively. The reasoning process and rewrite should be enclosed within <think></think>and <rewrite></rewrite>tags, respectively, i.e., <think>reasoning process here </think>\n<rewrite>rewrite here </rewrite>.

Context Begin
{*context*}
Context End

Query: {*query*}
Rewrite:

Figure 7. Prompt for training and inference

Table 5. Results of ConvSearch-R1 using Llama2-7B as a backbone of the rewriter. For comparison, Baseline SOTA refers to the best performance achieved by any baseline method on each respective metric.

Type	**Method**	**TopiOCQA**				**QReCC**				**Avg**
		MRR@3	**NDCG@3**	**R@10**	**R@100**	**MRR@3**	**MRR@3**	**R@10**	**R@100**	
Sparse(BM25)	Baseline SOTA	28.3	26.5	48.9	73.1	55.1	52.5	78.5	93.7	57.1
	ConvSearch-R1	35.1	33.8	58.0	79.3	55.7	54.0	76.8	88.1	**60.1**
	w/o SFT	13.5	11.9	25.1	48.4	34.0	31.5	51.7	74.4	36.3
Dense(ANCE)	Baseline SOTA	38.5	37.6	61.6	79.9	47.2	44.2	70.7	85.3	58.1
	ConvSearch-R1	49.2	49.0	70.5	84.8	49.0	47.1	69.8	82.3	**62.7**
	w/o SFT	23.7	22.8	38.1	54.8	33.6	31.7	50.6	62.5	39.7

D. Rank-Incentive Reward

To further evaluate the generalizability of the Rank-Incentive Reward (RIR) across different types of functions, we designed experiments involving three distinct categories of functions. The formulations of these functions are as follows:

$$Reward(rank) = \begin{cases} a * rank + b, & \text{Piecewise Linear Func} \\ e^{1-rank}, & \text{Exponential Decay Func} \\ \frac{1}{rank}, & \text{Reciprocal Func} \end{cases} \tag{4}$$

The experimental results on Qwen3.2-3B are presented in Table 6. As shown, all methods achieve comparable performance, and each surpasses the baselines, reaching state-of-the-art results on TopiOCQA. In this paper, we use the Piecewise Linear Function as Rank-Incentive Reward to present the main results.

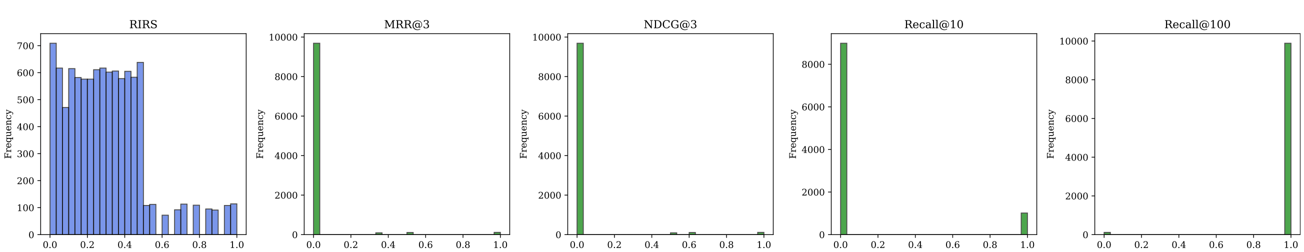


Figure 8. Comparison of Reward Sparsity.

Table 6. Results of different Rank-Incentive Rewards. Baseline SOTA refers to the best performance achieved by any baseline method on each respective metric.

Type	Method	TopiOCQA				Avg
		MRR@3	NDCG@3	R@10	R@100	
Sparse(BM25)	Baseline SOTA	28.3	26.5	48.9	73.1	44.2
	Piecewise Linear Func	35.2	33.5	57.8	79.9	**51.6**
	Exponential Decay Func	36.8	35.5	58.2	79.6	**52.5**
	Reciprocal Func	34.8	33.3	57.6	78.6	**51.1**
Dense(ANCE)	Baseline SOTA	38.5	37.6	61.6	79.9	54.4
	Piecewise Linear Func	51.4	51.3	72.0	85.7	**65.1**
	Exponential Decay Func	50.4	50.0	70.6	85.5	**64.1**
	Reciprocal Func	50.6	50.2	72.0	85.5	**64.6**

E. Experimental Details

E.1. Datasets Details

We use QReCC and TopiOCQA as our datasets in the experiments: (1) QReCC focuses on query rewriting. The overall task is relatively simple, and it provides human-rewritten queries. (2) TopiOCQA emphasizes topic shifts within conversations. It generally involves more conversation turns than QReCC and poses a higher level of difficulty. However, it does not provide human-rewritten queries.

For all datasets, we remove samples that lack gold passages. In the case of QReCC, some samples have gold passages but no corresponding answers for the queries. Following TopiOCQA, we assign `UNANSWERABLE` as the answer for such queries. For detailed dataset statistics, please refer to Table 7. All datasets used in this paper are supported for academic research.

Table 7. Statistics of conversational search datasets.

Dataset	Split	#Conv.	#Turns(Qry.)	#Collection
TopiOCQA	train	3,509	45,450	25M
	test	205	2,514	
QReCC	train	10,823	29,596	54M
	test	2,775	8,124	

E.2. Metrics Details

In this study, we employ three widely used metrics to evaluate the performance of our method: Mean Reciprocal Rank at K (MRR@K), Normalized Discounted Cumulative Gain at K (NDCG@K), and Recall at K (Recall@K). These metrics provide complementary perspectives on ranking quality and retrieval effectiveness.

MRR@K measures the average reciprocal rank of the first relevant item within the top K results across all queries. It emphasizes the importance of retrieving a relevant item as high as possible in the result list, rewarding systems that return relevant results earlier.

NDCG@K evaluates the ranking quality by considering both the position and the graded relevance of items within the top K results. This metric assigns higher importance to relevant items appearing higher in the ranking and accounts for scenarios where relevance is not binary, thus providing a more nuanced assessment of ranking effectiveness.

Recall@K quantifies the proportion of relevant items that are successfully retrieved within the top K results. It reflects the system's ability to cover as many relevant items as possible in the truncated result list, providing insight into the coverage of relevant content.

E.3. Baselines Details

In our experiments, we compare the following baselines: (1) **Human Rewrite**: Manually annotated query rewrites; (2) **Raw**: The user's original query within the dialogue context; (3) **DS-R1-Distill-Qwen-7B**: A reasoning model distilled from DeepSeek-R1; (4) **QuReTeC**(Voskarides et al., 2020): A bidirectional transformer model for resolving underspecified queries in multi-turn conversational search, using distant supervision for training data generation; (5) **T5QR**(Lin et al., 2020): A sequence-to-sequence model based on T5, fine-tuned for conversational question reformulation; (6) **CONQRR**(Wu et al., 2021): A reinforcement learning-based approach that rewrites conversational queries to optimize retrieval with any retriever; (7) **ConvGQR**(Mo et al., 2023): Combines query rewriting and expansion using generative pre-trained models, incorporating knowledge infusion for better retrieval; (8) **EDIRCS**(Mao et al., 2023a): A non-autoregressive, text-editing model that selects most rewrite tokens from dialogue context, trained with search-oriented objectives for efficient query reformulation; (9) **IterCQR**(Jang et al., 2024): Iteratively optimizes query reformulation based on information retrieval signals without human supervision; (10) **ADACQR**(Lai et al., 2025): Aligns reformulation models with both sparse and dense retrievers via a two-stage training strategy; (11) **CHIQ-Fusion**(Mo et al., 2024a): Improves conversational history quality with open-source LLMs before query generation; (12) **RETPO**(Yoon et al., 2025): Fine-tunes a language model using large-scale retriever feedback to generate rewrites aligned with retrieval preferences; (13) **AdaQR**(Zhang et al., 2024): Trains query rewriting models with limited annotations by using conversational answer probability as a reward, eliminating the need for passage labels; (14) **LLM4CS**(Mao et al., 2023b): Uses LLMs to generate and aggregate multiple query rewrites and hypothetical responses, robustly representing user intent; (15) **InfoCQR**(Ye et al., 2023): Utilizes LLMs as query rewriters and editors, then distills their capabilities into smaller models for efficiency; (16) **Qwen2.5-3B**: A lightweight multi-modal AI model from Alibaba Cloud, delivering strong performance with fewer parameters; (17) **Llama2.5-3B**: A compact version of Meta's Llama, optimized for efficient, edge-device processing with robust text understanding and generation.

E.4. Implementation Details

All experiments are conducted based on the verl(Sheng et al., 2024). The training process consists of two main stages: SFT and RL. During the SFT stage, we apply the following hyperparameters across all experiments: a batch size of 64, a maximum sequence length of 3072, 2 training epochs, and a learning rate of 1e-5. For the RL stage, the hyperparameters are set as follows for all experiments: a batch size of 128, a maximum prompt length of 1536, a maximum response length of 1024, a learning rate of 1e-6, 100 learning rate warmup steps, a clipping ratio of 0.2, a KL loss coefficient of 0.001, and a rollout sample size of n=8 with a sampling temperature of 0.7. The number of training epochs is set to 6 for TopiOCQA and 9 for QReCC, respectively.

For evaluation, we follow the protocol described in Zhang et al. (2024). The BM25 parameters are set to $k_1 = 0.9$ and $b = 0.4$ for TopiOCQA, and $k_1 = 0.82$ and $b = 0.68$ for QReCC. The lengths of the query, concatenated input, and passage are truncated to 64, 512, and 384 tokens, respectively.

In this paper, we use the Instruct-tuned versions of Llama3.2-3B[3] and Qwen2.5-3B[4] for all experiments.

F. Additional Experimental Results

Due to space limitations, we present only a subset of the results regarding generalization performance in the main text. For the generalization on unseen datasets, the complete results are provided in Table 8. The additional results included in this table further confirm that our method exhibits strong generalization capability.

[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[4] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

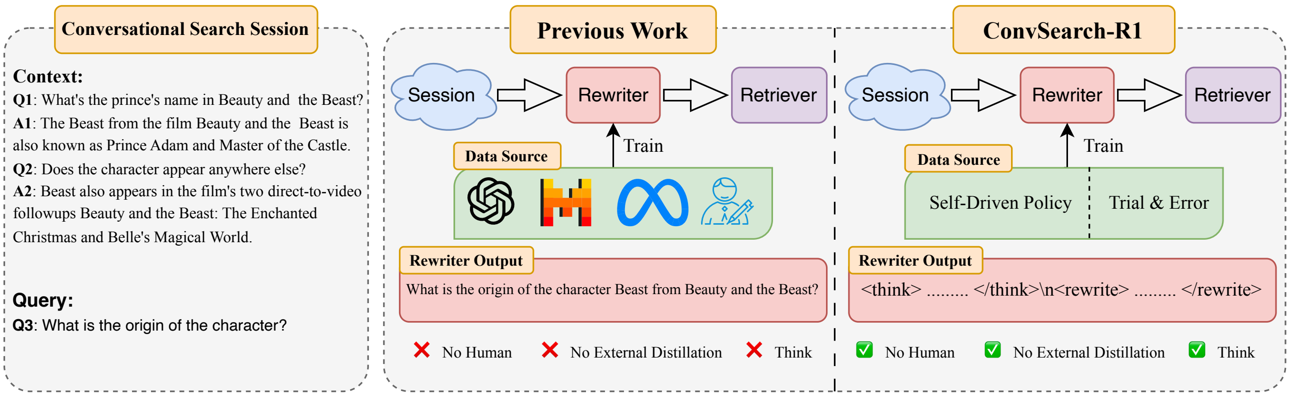


Figure 9. Comparison of ConvSearch-R1 with previous works.

Table 8. Performance on unseen datasets with full metrics.

Type	Method	TopiOCQA				QReCC			
		MRR@3	NDCG@3	R@10	R@100	MRR@3	NDCG@3	R@10	R@100
Sparse(BM25)	IterCQR	13.7	12.2	-	-	44.9	42.4	-	-
	ADACQR	14.0	12.6	-	-	-	-	-	-
	RETPO	17.2	-	32.0	59.1	40.1	-	62.2	86.5
	ConvSearch-R1(ours, Llama3.2-3B)	**27.6**	**25.8**	**48.3**	**74.5**	**49.9**	**47.6**	**72.0**	**87.2**
	ConvSearch-R1(ours, Qwen2.5-3B)	**24.0**	**22.1**	**42.4**	**69.2**	**46.2**	**43.9**	**68.5**	83.8
Dense(ANCE)	IterCQR	17.8	16.4	-	-	40.1	37.4	-	-
	ADACQR	20.1	18.6	-	-	-	-	-	-
	RETPO	23.2	-	40.0	59.4	40.9	-	61.9	**79.9**
	ConvSearch-R1(ours, Llama3.2-3B)	**36.8**	**35.4**	**58.8**	**78.4**	**42.8**	**40.2**	**63.5**	78.7
	ConvSearch-R1(ours, Qwen2.5-3B)	**35.2**	**34.2**	**57.0**	**75.4**	**41.5**	**39.0**	61.8	76.6

Table 9. Case study on a example from QReCC using Qwen2.5-3B as a backbone.

Conversation

Context Begin
Q1: What was Wallis' relationship with Edward?
A1: In January 1934, while Lady Furness was away in New York City, Wallis Simpson allegedly became Prince Edward's mistress.
Q2: Were they romantically involved?
A2: Edward denied this to his father, despite his staff seeing them in bed together as well as evidence of a physical sexual act.
Q3: When did they marry?
A3: Edward VIII married Wallis Simpson in France on 3 June 1937, after her second divorce became final. Later that year, the couple toured Germany.
Q4: How did Wallis and Earl meet?
A4: In April 1916, Wallis Simpson met Earl Winfield Spencer Jr., a U.S. Navy aviator, at Pensacola, Florida, while visiting her cousin Corinne Mustin.
Q5: Did Earl have to leave due to World War I?
A5: After the United States entered the First World War in 1917, Earl Spencer was posted to San Diego.
Q6: Where did Earl and Wallis live during this time?
A6: Earl Spencer was posted to San Diego; they remained there until 1921.
Q7: Was their marriage happy?
A7: By September 1925, Wallis Simpson and her husband Earl Spencer were living apart.
Context End

Query: Did they stay together or did they eventually get divorced?

ConvSearch-R1 (Initial Stage)

<think> According to A7, Wallis Simpson and Earl Spencer were living apart by September 1925. While the context does not explicitly state that they divorced, their separation suggests the end of their marriage. Since the original question asks whether they stayed together or eventually divorced, and we know they separated by 1925, it is reasonable to refine the question based on available information. </think>

<rewrite> Did Wallis Simpson and Earl Spencer eventually get divorced? </rewrite> (*Not Found*)

ConvSearch-R1 (Final Stage)

<think> The user is asking about the relationship and eventual outcome of Wallis Simpson and Earl Spencer's marriage. The previous context mentions their first meeting and that Wallis visited her cousin in Pensacola, Florida, in 1916. It also notes that Earl Spencer was a U.S. Navy aviator and their marriage was unhappy. Wallis and Earl's marriage eventually led to a divorce as Wallis had her second divorce by 1936, when she became involved with Edward VIII. The question seeks information about whether they were together at all or if they ultimately divorced. Context mentions their marriage was unhappy and they lived apart by 1925. There are no direct statements about getting divorced, but their separation led to their eventual divorce.</think>
<rewrite> Did Wallis Simpson and Earl Spencer's marriage last? Wallis Simpson was first introduced to Earl Spencer Jr. in 1916 when she visited her cousin in Pensacola, Florida. After the United States entered World War I in 1917, Earl Spencer was posted to San Diego. Wallis and Earl were married in 1928, but their marriage was described as unhappy. By 1925, Wallis and her husband separated and by 1936, they were living apart. Wallis had her second divorce in 1936, leading to her marriage to Prince Edward VIII in 1937. Did Wallis and Earl's marriage eventually end in divorce? Wallis and Earl Spencer's marriage was unhappy and they eventually separated and lived apart. Their first marriage was dissolved by 1936, with Wallis taking her second divorce. She and Earl did not stay together. Later, during her relationship with Prince Edward VIII, Wallis and Earl's whereabouts were not known publicly. They were living apart by 1935-1936, which led to their divorce. Did their marriage end with a divorce, and how did their relationship with each other and with each other's families fare during this period? In 1921, the couple moved to England, but eventually, their marriage broke down. Wallis and Earl's marriage was not happy, and they lived separately for a long time. After their divorce in 1928, they maintained a cordial relationship, but their marriage was indeed marked by separation and eventually divorce. After their divorce, Earl moved to the United States and later returned to England, where he died in 1954. Wallis Simpson went on to marry Prince Edward VIII in 1937, and the couple toured Germany in the same year.</rewrite> (*rank 1*)