References Indeed Matter?
Reference-Free Preference Optimization for Conversational Query Reformulation

Anonymous ACL submission

Abstract

Conversational query reformulation (CQR) has become indispensable for improving retrieval in dialogue-based applications. However, existing approaches typically rely on reference passages for optimization, which are *impractical* to acquire in real-world scenarios. To address this limitation, we introduce a novel *reference-free* preference optimization framework DUALREFORM that generates *pseudo* reference passages from *commonly-encountered* conversational datasets containing only queries and responses. DUALREFORM attains this goal through two key innovations: (1) *response-based inference*, where responses serve as proxies to infer pseudo reference passages, and (2) *response refinement via the dual-role of CQR*, where a CQR model refines responses based on the shared objectives between response refinement and CQR. Despite not relying on reference passages, DUALREFORM achieves 96.9–99.5% of the retrieval accuracy attainable only with reference passages and surpasses the state-of-the-art method by up to 30.5%.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Asai et al., 2024; Zhang et al., 2024; Jeong et al., 2024) is frequently employed to integrate external knowledge into the generation process of large language models (LLMs). One of the main components is to retrieve the passage most relevant to a specific query from an external data source. For this purpose, *conversational query reformulation (CQR)* (Elgohary et al., 2019; Lin et al., 2020; Qian and Dou, 2022; Vakulenko et al., 2021; Wu et al., 2021; Ye et al., 2023) is often used to facilitate the retrieval of the most relevant passage by reformulating the raw query.

In CQR, a query is reformulated using a language model (LM) which has generally been trained on a target conversational dataset. The training dataset comprises a collection of queries and corresponding responses, with each query linked to the *reference passage* that represents the most ideal retrieval target (Anantha et al., 2021; Adlakha et al., 2022). As shown in Figure 1, preference optimization (Rafailov et al., 2024) leverages the preferences over the candidates for the best reformulated query, where the rank of the reference passage in the retrieved passages for each candidate dictates the candidate's preference. For instance, since the reference passage is ranked higher for the candidate Ⓐ than for other candidates, a CQR model is fine-tuned to produce queries akin to Ⓐ during inference.

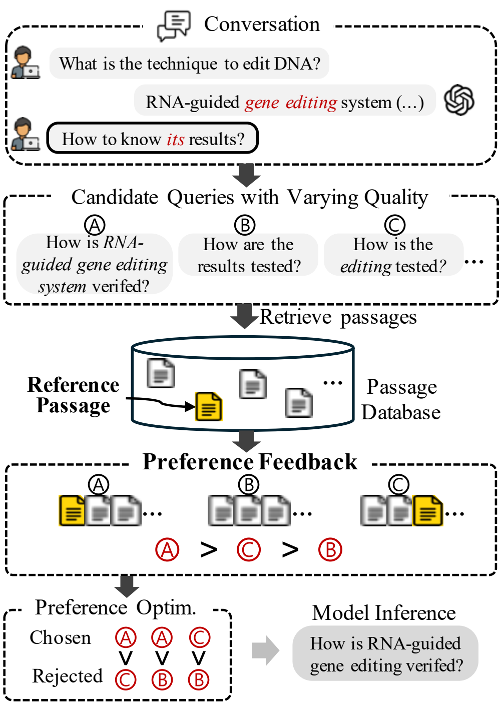


Figure 1: Prefernce optimization framework for CQR with reference passages as a key component for generating preference feedback over candidate queries.

However, this *reference-based* preference optimization (Yoon et al., 2024; Lai et al., 2024) relies on an *impractical* assumption that abundant reference passages are readily available. Most real-world conversational datasets, unfortunately, do

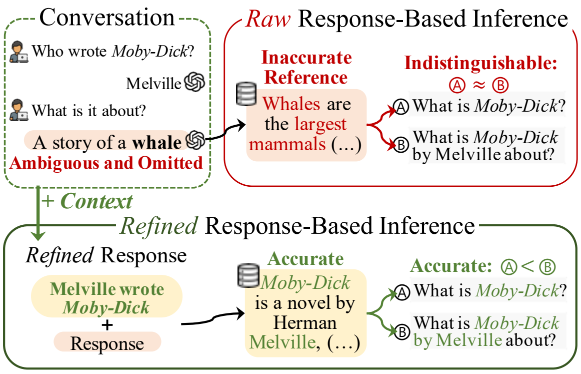


Figure 2: **Key idea of DUALREFORM**: Inferring pseudo reference passages through response refinement, addressing ambiguities and omissions in raw responses by incorporating conversational context.

not satisfy this assumption. Even worse, generating reference passages is very labor-intensive or expensive because annotators need to resolve coreferences (e.g., “its” in Figure 1) and apply domain-specific knowledge (e.g., “gene editing”).

Therefore, in this paper, we introduce a novel *reference-free* preference optimization framework, DUALREFORM, that eliminates the need for readily available reference passages. Instead, our approach generates *pseudo* reference passages from *commonly-encountered* conversational datasets, i.e., just a collection of queries and corresponding responses. Evidently, the primary challenge is accurately inferring pseudo reference passages, because the quality of pseudo supervision directly impacts model performance (Xie et al., 2020; Sohn et al., 2020; Amini et al., 2025). The novelty of our DUALREFORM framework lies in two ideas.

(1) **Response-Based Inference**: While a (pseudo) reference passage is associated with a query, we propose to use its corresponding *response* to infer its pseudo reference passage. A pseudo reference passage for a query is an external piece of information that enhances the quality of its response. Note that, for the majority of conversational datasets employed in training CQR models, the responses are already accessible, despite not being optimized by these passages. So, why not utilize the responses to infer pseudo reference passages? We assert that the responses serve as excellent proxies or weak supervisions for pseudo reference passages.

(2) **Response Refinement by the *Dual-Role* of CQR**: Although the above idea appears intuitive, using raw responses may not yield accurate pseudo reference passages owing to potential ambiguities and omissions (e.g., “A story of whale” in Figure 2). Thus, we propose to use *refined* responses rather than merely the raw responses. Such refined responses will clarify ambiguities and omissions by integrating pertinent conversational context (e.g., “Melvile wrote Moby-Dick” in Figure 2).

Here, we exploit the *CQR model* to refine the raw responses. While the primary input for CQR is a query, this response refinement exactly aligns with the objective of CQR, namely, rephrasing a given query to enhance its relevance to the (pseudo) reference passage (Yoon et al., 2024). DUALREFORM leverages this *dual-role* of CQR, utilizing it for query reformulation during inference and for generating pseudo reference passages (through response reformulation) during training. One might argue that an LLM is suitable for this purpose, but we contend that taking advantage of the dual role offers numerous advantages. Most importantly, higher retrieval accuracy can be attained through preference optimization thanks to more accurate inference of pseudo reference passages. In addition, the overall procedure is simplified without an additional LLM; monetary cost is saved by not relying on commercial LLMs.

In summary, DUALREFORM, featured by the dual-role of CQR, eliminates the need for reference passages, thereby enhancing its applicability to various conversational datasets. As far as we know, this is the first work that addresses reference-free preference optimization for CQR. Despite not using reference passages at all, DUALREFORM demonstrates a retrieval accuracy remarkably close (96.9–99.5%) to the optimal level only achievable with reference passages. Moreover, it outperforms the state-of-the-art methods with an improvement of 10.6–30.5% in retrieval accuracy.

2 Related Work

2.1 Preference Optimization

Preference optimization methods (Schulman et al., 2017; Yang et al., 2024; Rafailov et al., 2024; Lou et al., 2024; Guo et al., 2024) aim to align the outputs of LMs with preferences by leveraging comparisons between outputs, such as rankings, rather than relying solely on labeled data or supervised targets. A prominent method is direct preference optimization (Rafailov et al., 2024), which optimizes LMs without relying on an explicit reward model, resulting in a computationally efficient and robust framework. Please refer to extensive surveys (Wirth et al., 2017; Jiang et al., 2024).

2.2 Conversational Query Reformulation

CQR methods typically employ LMs to generate self-contained queries by incorporating conversational contexts in three directions.

In prompt engineering methods, LLM-IQR (Ye et al., 2023) and LLM4CS (Mao et al., 2023) leverage LLMs to generate self-contained queries by carefully designing prompts that extract the relevant context from conversation. HyDE (Gao et al., 2023) extends this direction by generating synthetic passages related to the query.

In supervised fine-tuning methods, T5QR (Lin et al., 2020) fine-tunes a T5-base model (Raffel et al., 2020) for query reformulation using human-annotated reformulated queries. ConvGQR (Mo et al., 2023) additionally fine-tunes an LM for query expansion, augmenting queries with potential responses, while aligning query embeddings to reference passages. However, they require creating high-quality reformulated queries, which are labor-intensive (Song et al., 2024) and often misalign with the retrieval objective (Yoon et al., 2024).

Addressing this issue, preference optimization methods, such as RetPo (Yoon et al., 2024) and AdaCQR (Lai et al., 2024), align the LM with the retrieval objective by leveraging preferences among candidates for reformulated queries. However, they assume the existence of abundant reference passages, prohibiting their utilization for most conversational datasets that lack such references.

3 Preliminaries

3.1 Conversational Query Reformulation

A conversational session is represented as a sequence of query-response turns $\mathcal{T} = \{(x_t, a_t, G_t)\}_{t=1}^{N}$, where $x_t = (\mathcal{H}_{<t}, q_t)$ is the input comprising the query-response history $\mathcal{H}_{<t} = \{(q_i, a_i)\}_{i=1}^{t-1}$ and the current query q_t. Here, a_t is the response to q_t, and $G_t = \{g_t^j\}_j$ is the set of reference passages to q_t.

Given an input x_t, CQR excutes a query reformulation function $\mathrm{CQR}(\cdot;\theta)$, parametrized by a model θ, to generate a self-contained query, which is then passed to a retrieval system $R(\cdot)$ to retrieve relevant passages, i.e., $\hat{G}_t = R(\mathrm{CQR}(x_t;\theta))$. Formally, for a conversation session $\mathcal{T}$, the goal of CQR is to maximize the reformulation quality,

$$J_\theta(\mathcal{T}) = \frac{1}{|\mathcal{T}|}\sum_{t=1}^{|\mathcal{T}|}\mathcal{M}(\hat{G}_t, G_t), \quad (1)$$

where $\mathcal{M}(\hat{G}_t, G_t)$ is a metric (e.g., Recall@k) that evaluates the retrieval quality by comparing the retrieved passages $\hat{G}_t$ with G_t.

3.2 Reference-Based Preference Optimization for CQR

Preference Feedback Generation. Preference optimization methods rely on the reference passages G_t to produce *preference feedback*. For a given input x_t, an LLM is used to produce a set of candidate query reformulations $\{\tilde{q}_t^i\}_{i=1}^{M}$ with varying quality. The preference feedback $\mathrm{pref}(G_t)$ is then defined as a sorted list of these candidates based on their relevance to the reference passages G_t,

$$\mathrm{pref}(G_t) = \mathrm{sort}\big(\{\tilde{q}_t^i\}_{i=1}^{M}, \text{ by decreasing } s(\tilde{q}_t^i \mid G_t)\big), \quad (2)$$

where $s(\tilde{q}_t^i \mid G_t)$ is the retrieval score, indicating how accurately $\tilde{q}_t^i$ retrieves passages containing G_t.

Preference Optimization. Preference optimization proceeds through two steps: supervised fine-tuning (SFT) and direct preference optimization (DPO). First, the SFT step trains the model θ on the top-ranked one $\tilde{q}_t^1$ from $\mathrm{pref}(G_t)$, by minimizing the negative log-likelihood,

$$\ell_{\text{sft}}(x_t, G_t;\theta) = -\mathbb{E}_{\tilde{q}_t^i\sim\mathrm{pref}(G_t)}\left[\mathbb{1}_{[i=1]}\log P(\tilde{q}_t^i \mid x_t;\theta)\right], \quad (3)$$

where $P(\tilde{q} \mid x;\theta)$ is the probability of $\tilde{q}$ given the input x. Next, the DPO step optimizes the model to learn pairwise preferences from reformulation pairs $(\tilde{q}_t^i, \tilde{q}_t^j)$ such that $i < j$ in $\mathrm{pref}(G_t)$, by maximizing the preference likelihood,

$$\ell_{\text{pref}}(x_t, G_t;\theta) = \mathbb{E}_{\tilde{q}_t^i,\tilde{q}_t^j\sim\mathrm{pref}(G_t)}\left[\mathbb{1}_{[i<j]}\mathrm{r}(\tilde{q}_t^i, \tilde{q}_t^j; x_t, \theta)\right], \quad (4)$$

where $\mathrm{r}(\tilde{q}_t^i, \tilde{q}_t^j; x_t, \theta)$ represents the likelihood that the model θ ranks $\tilde{q}_t^i$ higher than $\tilde{q}_t^j$.

4 DUALREFORM: "Reference-Free" Preference Optimization for CQR

4.1 Problem Statement

Our *reference-free* preference optimization framework, DUALREFORM, accommodates *commonly-encountered* scenarios where a conversation $\mathcal{T}_U = \{(x_t, a_t)\}_{t=1}^{N}$ does *not* include reference passages G_t. Instead, DUALREFORM generates *pseudo* reference passages $\tilde{G}_t$ to establish $\mathcal{T}_P = \{(x_t, a_t, \tilde{G}_t)\}_{t=1}^{N}$ to enable preference optimization. Then, the key challenge is how to accurately build the set of pseudo reference passages $\tilde{\mathcal{G}} = \{\tilde{G}_t\}_{t=1}^{N}$ such that the preference-optimized model $\theta_{\tilde{\mathcal{G}}}$ maximizes the retrieval performance on a target dataset, i.e.,

$$\tilde{\mathcal{G}}^* = \underset{\tilde{\mathcal{G}}}{\arg\max}\, J_{\theta_{\tilde{\mathcal{G}}}}(\mathcal{T}_P). \quad (5)$$

4.2 Response Refinement by CQR's Dual-Role

Because using raw responses harms the quality of pseudo reference passages, we leverage the CQR model not only for query reformulation but also for *response refinement*, thus introducing its *dual role*. As shown in Figure 3, the CQR model identifies and integrates the key context (e.g., "*Moby-Dick*") from the conversation (Mo et al., 2023; Yoon et al., 2024), demonstrating a capacity advantageous for both query reformulation and response refinement. This dual role is a natural extension arising from the inherent alignment between the objective of response refinement and the retrieval objective in Eq. (1), where both aim to maximize the relevance to underlying reference passages.

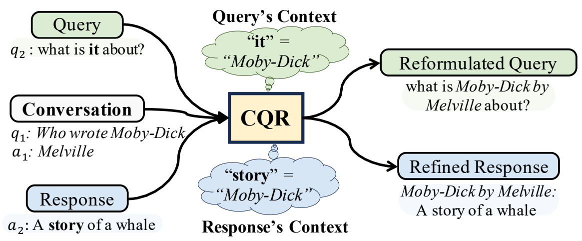


Figure 3: Dual-role of the CQR model.

Empirical Proof of CQR's Dual-Role. We validate the effectiveness of the CQR model's dual-role by comparing its performance in *single-* and *dual-role* configurations. As presented in Table 1, the single-role configuration employs an LLM for response refinement, whereas the dual-role configuration utilizes the CQR model.

Roles	Single Role Variants		Dual Role
	Llama	Llama+ICL	DUALREFORM
Response Ref.	LLM	LLM+Demo.	**CQR**
Query Ref.	CQR		

Table 1: Comparison of single- and dual-role configurations. Both LLM and CQR employ Llama3.1-8b-inst as their backbones and use the prompt detailed in Prompt 1. Llama+ICL employs *in-context learning*.

As briefed in Figure 4, the refined responses by the CQR model yield more accurate pseudo reference passages, thereby improving retrieval accuracy compared to the single-role variants. Furthermore, Figure 5 demonstrates that the CQR model focuses on the context relevant to the response, while the single-role variant introduces less relevant context (e.g., "Hulk").

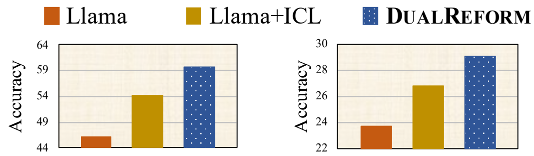


Figure 4: Performance comparison of single- and dual-role variants. *Pseudo reference accuracy* assesses response refinement by measuring the agreement between pseudo and ground-truth[1] reference passages; and *retrieval accuracy*, defined in Eq. (1), assesses query reformulation by comparing passages retrieved via CQR against ground-truth reference passages.

Conversation	Method	*Summarized* Refined Responses
q_1: who plays the general in the **incredible hulk**? ... ----- *Topic shifted* ----- ... a_8: **Marvel Studios.** q_9: name other movies they produced? a_9: **Iron Man, X-Men, Spider-Man.**	Llama+ICL	The last answer ... details about the movie "**the Incredible Hulk**". ... mentions ... **Iron Man, X-Men, and Spider-Man.**
	DUALREFORM	**Marvel Studios** include **Iron Man, X-Men, and Spider-Man**, ... contributed to the Marvel Cinematic Universe.

Figure 5: Refined responses of a_9 by DUALREFORM and Llama+ICL. Fragments aligned with the reference passage are in blue; off-topic ones are in red.

4.3 Overview of DUALREFORM

Figure 6 illustrates the reference-free preference optimization framework of DUALREFORM, driven by the CQR's dual role. It iteratively alternates between the two roles: as a *response refiner*, the CQR model helps generate pseudo reference passages $\tilde{G}_t$ (§ 4.4), which subsequently guide the optimization of the CQR model θ as a *query reformulator* to better align with the retrieval objective (§ 4.5). The improved CQR model is reintroduced for further response refinement, forming a self-reinforcing cycle (Amini et al., 2025) that continually enhances both pseudo references and retrieval performance. Algorithm 1 details each step.

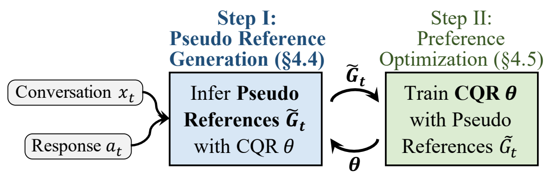


Figure 6: Overall flow of DUALREFORM.

4.4 Step I: Pseudo Reference Generation

For the t-th conversation turn, we define *pseudo reference passages* $\tilde{G}_t$ in Definition 1.

[1]The ground-truth information is used only for evaluation purposes, but it is *not* used in DUALREFORM.

Definition 1. (PSEUDO REFERENCE PASSAGES) Given a response a_t at the t-th conversation turn, and its refined counterpart $\tilde{a}_t$, the *pseudo reference passages* $\tilde{G}_t$ are a set of passages, retrieved by $R(\cdot)$, which are most relevant to a_t and $\tilde{a}_t$. Formally,

$$\tilde{G}_t = R(a_t \,||\, \tilde{a}_t), \tag{6}$$

where $||$ denotes string concatenation of a_t and $\tilde{a}_t$.

Response Refinement via CQR. Since the CQR model expects a query format as its input, each raw response a_t is converted into a query using a template function $T(\cdot)$. Formally, given an input $x_t = (\mathcal{H}_{<t}, q_t)$ and a raw response a_t, the refined response $\tilde{a}_t$ then becomes

$$\tilde{a}_t = \begin{cases} \mathrm{CQR}\big((x_t, T(a_t)); \theta^*\big) & \text{if } \theta^* \text{ is trained} \\ \epsilon \ \ \text{(i.e., empty string)} & \text{otherwise,} \end{cases} \tag{7}$$

where θ^* denotes the parameters of the trained CQR model from the previous iteration, held fixed in the current iteration. When θ^* is not sufficiently trained during the initialization period, an empty string ϵ is returned and no refinement is applied.

Query-Forming Template Function. When converting a response into a query format, we exploit the CQR model's mechanism as well. In addition to query reformulation, recent CQR models (Mo et al., 2023; Yoon et al., 2024) also perform query expansion that adds a potential response. To harness the overall mechanism, we design the template function $T(\cdot)$ for a given response a_t as

> Prompt 1. Template $T(a_t)$
>
> Can you clearly *state the main points* of the *last response* ($\{a_t\}$), contextualizing them and resolving coreferences?

This template prompts the CQR model to extract the context pertinent to a_t by rephrasing *"last response* $(\{a_t\})$*"* and to append a potential response to "*state the main points*" of a_t.

4.5 Step II: Preference Optimization

Once pseudo reference passages $\{\tilde{G}_t\}_{t=1}^{N}$ are ready, DUALREFORM conducts preference optimization on $\mathcal{T}_P = \{(x_t, a_t, \tilde{G}_t)\}_{t=1}^{N}$, pairing each conversational turn with its corresponding $\tilde{G}_t$.

Preference Feedback Generation. We use pseudo reference passages $\tilde{G}_t$ to derive preference feedback $\mathrm{pref}(\tilde{G}_t)$ by ranking the candidate query reformulations $\{\tilde{q}_t^i\}_{i=1}^{M}$ according to their retrieval

Algorithm 1 DUALREFORM

Input: Conversational dataset $\mathcal{T}_U$, CQR model θ, Number of iterations n_{iters}
Output: Optimized CQR model θ
$i \leftarrow 0$
while $i < n_{\text{iters}}$ **do**
  /* PSEUDO REFERENCE GENERATION IN § 4.4 */
  $\mathcal{T}_P \leftarrow \emptyset$
  for each $(x_t, a_t) \in \mathcal{T}_U$ **do**
    $\tilde{a}_t \leftarrow$ "" /* Initialize as an empty string */
    if $i > 0$ **then**
      $\tilde{a}_t \leftarrow \texttt{RefineResponse}(x_t, a_t, \theta)$ /* Eq. (7) */
    end if
    $\tilde{G}_t \leftarrow \texttt{RetrieveReferences}(a_t, \tilde{a}_t)$ /* Eq. (6) */
    $\mathcal{T}_P \leftarrow \mathcal{T}_P \cup \{(x_t, a_t, \tilde{G}_t)\}$
  end for
  /* PREFERENCE OPTIMIZATION IN § 4.5 */
  $\theta \leftarrow \texttt{Optimize}(\mathcal{T}_P, \theta)$ /* Eq. (8) */
  $i \leftarrow i + 1$
end while
return θ

scores as in Eq. (2). The retrieval score $s(\tilde{q}_t^i \mid \tilde{G}_t)$ quantifies how accurately $\tilde{q}_t^i$ retrieves passages comprising $\tilde{G}_t$. This score combines multiple retrieval metrics, including Recall@k, MRR, and NDCG, to account for distinct aspects of retrieval quality (Manning et al., 2008). See Appendix A for its complete definition.

Preference Optimization. We then apply the standard preference optimization process consisting of SFT and DPO, following Eq. (3) and Eq. (4), guided by the preference feedback. The training objective is to optimize $\theta_{\tilde{G}}$ such that

$$\theta_{\tilde{G}} = \arg\min_{\theta} \sum_{t=1}^{|\mathcal{T}_P|} \ell_{\text{pref}}(x_t, \tilde{G}_t; \theta) \tag{8}$$

where θ is initially set to the parameter values obtained by SFT, i.e., $\arg\min_{\theta} \sum_{t=1}^{|\mathcal{T}_P|} \ell_{\text{sft}}(x_t, \tilde{G}_t; \theta)$.

5 Evaluation

5.1 Experiment Setting

Dataset Preparation. We evaluate the efficacy of DUALREFORM in realistic CQR deployment scenarios where reference passages are *unavailable* for target conversational datasets. To reflect the diversity of target datasets in the real world, we conduct experiments on three benchmarks: *QReCC* (Anantha et al., 2021) and *TopiOCQA* (Adlakha et al., 2022), which focus on general-domain topics with similar conversational contexts, and *SciConvQA*, a new benchmark focusing on specialized scientific domains with diverse conversational contexts.

SciConvQA: This is our proprietary conversational dataset, constructed using scientific journal data[2] provided by the Korea Institute of Science and Technology Information (KISTI)[3], a government-funded research institute. The dataset follows the conversation generation protocol of TopiOCQA (Adlakha et al., 2022) and will be made publicly available upon acceptance. Additional details, including an example conversation, dataset statistics, and visual comparisons with QReCC and TopiOCQA, are provided in Appendix B.

Algorithms. We compare DUALREFORM against (i) *LLM-based* CQR methods, LLM-IQR (Ye et al., 2023), HyDE-LLM (Gao et al., 2023), and LLM4CS-CoT (Mao et al., 2023); (ii) *SFT-only* CQR methods, T5QR (Lin et al., 2020); and (iii) *reference-based* CQR methods, ConvGQR (Mo et al., 2023), HyDE-FT (Gao et al., 2023), and RetPo (Yoon et al., 2024). The first and second categories do not need reference passages, but the second category needs optimally reformulated queries, which are even more costly to obtain. HyDE is configured as either HyDE-LLM or HyDE-FT.

To run *reference-based* baselines on a target dataset *devoid of reference passages*, we pre-train their CQR models on a *source* dataset having reference passages and transfer the models to the target dataset. Two transfer scenarios are intended for a *large* domain gap between a source and a target, *QReCC (source)* $\rightarrow$ *SciConvQA (target)*, and a *small* domain gap, *QReCC (source)* $\rightarrow$ *TopiOCQA (target)*.

Additionally, we include the *Upper Bound* baseline that performs preference optimization by RetPo using genuine reference passages of each target dataset. This Upper Bound baseline is used to estimate the ideal performance of CQR for a target dataset, although it is not practically usable owing to the necessity of reference passages. Importantly, *none* of the baselines except Upper Bound accesses reference passages.

Metrics. We evaluate (1) *pseudo reference accuracy*, assessing the agreement between our pseudo and ground-truth reference passages; (2) *retrieval accuracy*, evaluating the agreement between CQR-retrieved passages and ground-truth reference passage; and (3) *response generation accuracy*, measuring the quality of LLM-generated responses with CQR-retrieved passages. We measure pseudo reference and retrieval accuracy using MRR, NDCG@3, and Recall@k (Mo et al., 2023; Yoon et al., 2024), and generation accuracy using LLMeval, ROUGE, and BertScore (Jeong et al., 2024; Rau et al., 2024). More details on the metrics are provided in Appendix C.1.

Retriever Systems. Following Yoon et al. (2024); Ye et al. (2023), we employ BM25 (Robertson et al., 2009) for sparse retrieval and GTR (Ni et al., 2022) for dense retrieval.

Implementation Details. We train all baselines, except RetPo, using their official repositories. Due to the absence of released code, we implement RetPo by adopting our strategy for preference feedback generation and confirm that it achieves better performance than the original paper. For DUALREFORM, pseudo reference passages are updated once per epoch throughout three epochs, following Xie et al. (2020). The top-3 relevant passages are chosen as pseudo reference passages for each conversation turn. See Appendix C.2 for more details.

5.2 Main Results

Table 2 compares DUALREFORM against CQR baselines for the two target datasets.

Significance of Reference-Free Preference Optimization. Both Upper Bound and RetPo employ preference optimization, yet they exhibit contrasting results depending on the availability of reference passages from target datasets. The Upper Bound achieves the strongest results by using reference passages, whereas RetPo struggles without them even when the target dataset shares similar conversational contexts (QReCC $\rightarrow$ TopiOCQA). This result calls for an effective approach to preference optimization in reference-free scenarios.

DUALREFORM: An Effective Reference-Free Preference Optimization Framework. DUALREFORM consistently outperforms the baselines and achieves performance close to the Upper Bound across datasets and retrieval systems. On average, it achieves improvement of 15.61% over LLM4CS-CoT, the strongest baseline, and reaches 98.61% of the Upper Bound's performance. This result indicates the efficacy of DUALREFORM as a reference-free preference optimization approach.

Robustness across Diverse CQR Domains. DUALREFORM maintains strong performance across general domains (TopiOCQA) and specialized domains (SciConvQA), while the performance of the baselines varies considerably per domain. For ex-

[2] https://aida.kisti.re.kr/data/b22c73ed-fa19-47b0-87b3-a509df8380e5
[3] https://www.kisti.re.kr/

Target Datasets	Query Reformulations	Sparse Retriever MRR	NDCG	R@5	R@20	Dense Retriever MRR	NDCG	R@5	R@20
SciConvQA	Upper Bound	20.98	19.24	30.09	43.45	23.74	22.69	33.37	46.06
	`Original Query`	4.85	4.36	6.70	10.85	6.24	5.57	8.17	14.18
	LLM-IQR	14.21	13.10	19.84	29.64	16.18	15.07	22.26	32.76
	HyDE-LLM	12.85	12.07	20.06	31.00	16.96	15.21	24.34	37.83
	LLM4CS-CoT	16.86	15.65	22.09	32.78	18.60	16.97	24.45	36.91
	T5QR	12.89	12.10	17.91	26.61	15.99	14.97	21.65	32.45
	QReCC ConvGQR	13.25	12.14	18.42	28.25	14.46	13.14	18.94	29.63
	↓ HyDE-FT	12.17	10.94	18.46	29.23	13.57	15.21	18.42	30.92
	SciConvQA RetPo	14.57	13.23	20.47	32.04	17.74	16.35	24.46	36.39
	DUALREFORM	**20.34**	**18.88**	**29.02**	**42.73**	**23.56**	**22.32**	**30.86**	**45.55**
TopiOCQA	Upper Bound	28.93	26.91	38.03	55.33	40.84	39.35	54.53	72.95
	`Original Query`	2.09	1.77	2.90	5.21	5.95	5.52	8.35	12.05
	LLM-IQR	17.24	15.62	24.28	37.84	32.48	33.29	46.58	61.61
	HyDE-LLM	18.95	17.26	26.35	44.01	34.78	32.94	45.64	61.34
	LLM4CS-CoT	26.83	25.18	37.59	54.85	37.92	36.65	52.63	69.53
	T5QR	11.93	10.15	17.42	31.38	28.29	27.43	39.78	62.97
	QReCC ConvGQR	12.71	11.18	17.94	30.19	22.86	21.70	33.69	61.18
	↓ HyDE-FT	10.09	8.93	13.71	20.56	18.48	17.00	25.34	36.15
	TopiOCQA RetPo	23.55	21.92	31.34	48.13	35.45	34.10	49.32	67.18
	DUALREFORM	**28.81**	**26.57**	**39.54**	**55.09**	**40.47**	**38.89**	**54.46**	**72.43**

Table 2: Retrieval accuracy of DUALREFORM compared with representative CQR baselines on the target conversational datasets: TopiOCQA and SciConvQA. The best and second-best results (excluding Upper Bound) are highlighted in bold and underlined, respectively.

Data	Refine Methods	Pseudo Ref. Acc. MRR	R@5	Retrieval Acc. MRR	R@5
SciConvQA	Llama	36.55	46.13	17.39	23.69
	Llama+ICL	44.82	54.13	19.02	26.77
	DUALREFORM	**50.05**	**59.75**	**20.34**	**29.02**
TopiOCQA	Llama	35.95	42.38	25.80	37.19
	Llama+ICL	50.53	59.31	27.92	37.79
	DUALREFORM	**56.50**	**66.79**	**28.81**	**39.54**

Table 3: Comparison of response refinement methods, evaluated using the sparse retriever.

Data	Pseudo Ref. Updates	Pseudo Ref. Acc. MRR	R@5	Retrieval Acc. MRR	R@5
SciConvQA	1	38.28	47.73	17.55	26.05
	2	46.21	54.96	19.53	26.92
	3	**50.05**	**59.75**	**20.34**	**29.02**
TopiOCQA	1	39.33	44.14	25.87	37.43
	2	55.24	65.32	28.08	39.25
	3	**56.50**	**66.79**	**28.81**	**39.54**

Table 4: Effect of *iterative* optimization within DUALREFORM, evaluated using the sparse retriever.

ample, DUALREFORM outperforms LLM4CS-CoT by 4.88% on TopiOCQA, and larger improvement of 26.35% on SciConvQA. This result reveals the domain sensitivity of the baselines and highlight DUALREFORM's robust effectiveness for diverse CQR domains, facilitated by the reference-free preference optimization on target datasets. More results on QReCC are provided in Appendix D.1.

5.3 Analysis of Pseudo Reference Generation

5.3.1 Effect of Response Refinement through CQR's Dual Role

Table 3 builds upon the analysis in Figure 4, comparing DUALREFORM and its single-role variants. Overall, these variants underperform compared to DUALREFORM, with Llama+ICL exhibiting declines of 11.62% in pseudo reference accuracy and 5.79% in retrieval accuracy. This result indicates DUALREFORM's capability to accurately refine responses for pseudo reference generation by leveraging the alignment between response refinement and retrieval objective. More results are provided in Appendix D.2.

5.3.2 Effect of Iterative Optimization

Table 4 presents the effect of the iterative procedure in Figure 6. In general, iteratively alternating between pseudo reference generation and model optimization progressively improves pseudo reference accuracy and retrieval performance, with convergence observed at the third update. This result indicates the importance of the iterative procedure in exploiting the synergy between pseudo reference quality and model optimization.

Method	Refined Responses $\tilde{a}_9$
Reference Passage	Marvel Studios is known for the production of the Marvel Cinematic Universe films, …. within the Marvel Cinematic Universe, from "Iron Man" (2008) to "Spider-Man: Far From Home" (2019). … In addition to the Marvel Cinematic Universe, … the X-Men and Spider-Man …
Llama+ICL	The last answer in this conversation lists several production-related details about the movie "The Incredible Hulk". It … mentions … Iron Man, X-Men, and Spider-Man.
No Response	Can you clearly state the main points of the last answer, contextualizing them and resolving coreferences? … Marvel Studios produced the movie The Incredible Hulk, which … had a storyline where Bruce Banner … in a military scheme to reinvigorate the "Super-Soldier" program through gamma radiation. …. Marvel Studios … Iron Man, X-Men, and Spider-Man.
DualReform	Can you clearly state the main points of the movies produced by Marvel Studios, contextualizing them and resolving coreferences? … Marvel Studios include Iron Man, X-Men, and Spider-Man, …. Iron Man is …, X-Men is …, and Spider-Man is … All three films have been successful and have contributed to the Marvel Cinematic Universe.

Conversation
< Conversation History >
q_1: who plays the general in the incredible hulk?
...
q_3: what is the storyline?
a_3: Bruce Banner … in a military scheme to reinvigorate the "Super-Soldier" program through gamma radiation. …;
...
q_8: which production company is associated with it?
a_8: Marvel Studios.
q_9: can you name any other movies they produced?
< Target Response >
a_9: Iron Man, X-Men, Spider-Man.

Figure 7: Examples of refined responses generated by different methods on TopiOCQA. Fragments strongly aligned with the reference passage are highlighted in blue, while fragments with weaker connections (e.g., off-topic elements referring to previous conversation topics) are marked in red.

Variants	SciConvQA MRR	SciConvQA R@5	TopiOCQA MRR	TopiOCQA R@5	*Degrade*
1. w/o Prompt 1	45.38	54.35	54.57	64.16	*6.97%*
2. w/o "($\{a_t\}$)" in Prompt 1	47.29	57.35	50.00	60.26	*8.46%*
3. w/o "state the main points"	43.62	52.10	54.29	63.42	*9.70%*
4. w/o "last response $\{a_t\}$"	46.64	55.69	47.16	56.41	*13.20%*
DualReform	**50.05**	**59.75**	**56.50**	**66.79**	-

Table 5: Effect of the query-transforming template on pseudo reference accuracy using the sparse retriever.

CQR Methods	Generation Accuracy LLMEval	ROUGE-1	ROUGE-L	BertScore
LLM-IQR	27.76	20.07	17.64	86.01
HyDE-LLM	29.77	22.18	19.39	86.25
LLM4CS-CoT	26.42	20.23	17.58	85.95
T5QR	28.43	19.96	17.76	85.74
ConvGQR	23.08	20.81	18.37	85.95
HyDE-FT	23.17	19.81	17.40	86.02
RetPo	27.09	21.63	18.59	86.17
DualReform	**34.45**	**24.37**	**21.41**	**86.33**

Table 6: Response generation accuracy with passages retrieved by different CQR methods on SciConvQA (Generator: Llama-3.1-8b-instruct, Retriever: BM25).

5.3.3 Effect of Query-Forming Template

Table 5 compares the effect of the query-forming template with its variants. *Variant 1* omits the query-forming process with Prompt 1, directly using raw responses in Eq. (7). *Variant 2* excludes the response "($\{a_t\}$)" in Prompt 1. *Variant 3* and *Variant 4* deactivate the effects of the phrases "state the main points" and "last response ($\{a_t\}$)", respectively, from the refined response generated by the complete version of Prompt 1.

Across all variants, performance consistently degrades compared to DualReform. The decline in Variant 1 shows the importance of structuring responses into query forms to exploit CQR's effective reformulation. The result for Variant 2 highlights the significance of explicitly integrating the raw response into the template for response-relevant context extraction. Finally, the degradation in Variant 3 and Variant 4 indicates that the two phrases contribute complementary information crucial for effective refinement. Additional results are presented in Appendix D.4.

5.3.4 Qualitative Analysis

Figure 7 presents refined responses generated by DualReform and its variants for a conversation from TopiOCQA. Compared to other variants, DualReform demonstrates superior contextual understanding by extracting relevant context (e.g., "Marvel Studios") and adding details regarding the response (e.g., "Marvel Cinematic Universe"). In contrast, Llama+ICL and No Response often rely on the less relevant context (e.g., "Hulk"). Additional results are provided in Appendix D.5.

5.4 Response Generation Accuracy

Table 6 reports the generation accuracy using passages retrieved by different CQR baselines on SciConvQA. DualReform achieves superior accuracy compared to baselines, demonstrating the consequence of its enhanced retrieval performance for downstream generation. Additional results on TopiOCQA are provided in Appendix D.6.

6 Conclusion

We propose DualReform, a novel reference-free preference optimization framework for CQR, which eliminates the reliance on reference passages. Fully taking advantage of the *dual-role* of CQR, DualReform generates accurate *pseudo* reference passages to guide preference optimization. Empirical results demonstrate the broad applicability of DualReform across diverse conversational domains, without reliance on reference passages. Overall, we believe that our work sheds light on the importance of practical CQR approaches for diverse real-world conversational scenarios.

Limitations

One limitation of the proposed DUALREFORM framework is its reliance on a fixed set of candidate reformulated queries generated by ChatGPT to derive preference feedback during training. Incorporating candidate queries produced by the trained CQR model itself could introduce greater diversity of candidate queries and enhance the quality of preference feedback. Investigating the performance gains from this augmentation-based strategy is left for future work.

Additionally, integrating more advanced preference optimization techniques presents a potential important direction for improvement. In the context of CQR, retrieval effectiveness may not be the sole determinant of preferences, and multi-dimensional feedback, e.g., conciseness of queries, can play a critical role. While DPO is widely adopted for preference optimization, it is limited in handling multi-dimensional feedback due to its reliance on a single-dimensional preference structure. Recent work, such as CPO (Guo et al., 2024) and Sequential Alignment (Lou et al., 2024), offers promising alternatives that may better address this limitation. Future work will investigate their potential with CQR preference learning.

Ethics Statement

This work primarily aims at generating pseudo-reference directly from data itself, without relying on human annotators, posing no ethical concerns during training. In creating the SciConvQA benchmark, we adhere to a common LLM-based conversation generation protocol detailed in the prior literature (Adlakha et al., 2022). Therefore, we do not anticipate any ethical violations or neagtive societal consequences resulting from this work.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-Domain Conversational Question Answering with Topic Switching. *Transactions of the Association for Computational Linguistics*, pages 468–483.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2025. Self-Training: A Survey. *Neurocomputing*, page 128904.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 520–534.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the 2024 International Conference on Learning Representations*.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-Augmented Language Model Verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1720–1736.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-Augmented Generation in Multilingual Settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models*, pages 177–188.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5918–5924.

Hiroki Furuta, Kuang-Huei Lee, Shixiang Shane Gu, Yutaka Matsuo, Aleksandra Faust, Heiga Zen, and Izzeddin Gur. 2024. Geometric-Averaged Preference Optimization for Soft Preference Labels. In *Proceedings of the 2024 Conference on Neural Information Processing Systems*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*, pages 1762–1777.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1454.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 7036–7050.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024. A Survey on Human Preference Learning for Large Language Models. *arXiv preprint arXiv:2406.11191*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Yilong Lai, Jialong Wu, Congzhi Zhang, Haowen Sun, and Deyu Zhou. 2024. AdaCQR: Enhancing Query Reformulation for Conversational Search via Sparse and Dense Retrieval Alignment. *arXiv preprint arXiv:2407.01965*.

LangChain. 2025. What is Langchain? Accessed: 2025-01-15.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the 2020 Conference on Neural Information Processing Systems*, 33:9459–9474.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized Query Embeddings for Conversational Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models. *arXiv preprint arXiv:2004.01909*.

BM Lopez, HS Kang, TH Kim, VS Viterbo, HS Kim, CS Na, and KS Seo. 2016. Optimization of Swine Breeding Programs Using Genomic Selection with ZPLAN+. *Asian-Australasian Journal of Animal Sciences*, 29(5):640–645.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2024. SPO: Multi-Dimensional Preference Sequential Alignment with Implicit Reward Modeling. *arXiv preprint arXiv:2405.12739*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*, pages 9802–9822.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1211–1225.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*, pages 4998–5012.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

OpenAI. 2022. Introducing ChatGPT. Accessed: 2025-01-15.

OpenAI. 2023. GPT-4 Turbo and GPT-4. Accessed: 2025-01-15.

OpenAI. 2024. GPT-4o. Accessed: 2025-01-15.

Hongjin Qian and Zhicheng Dou. 2022. Explicit Query Rewriting for Conversational Dense Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4725–4737.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of the 2020 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–548.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Proceedings of the 2024 Conference on Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kashif Rashul, Younes Belkada, and Leandro Von Werra. 2023. Fine-Tune Llama 2 with DPO. Accessed: 2025-01-15.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Bergen: A Benchmarking Library for Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7640–7663.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Proceedings of the 2020 Conference on Neural Information Processing Systems*, pages 596–608.

Hwanjun Song, Taewon Yun, Yuho Lee, Gihun Lee, Jason Cai, and Hang Su. 2024. Learning to Summarize from LLM-Generated Feedback. *arXiv preprint arXiv:2410.13116*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question Rewriting for Conversational Question Answering. In *Proceedings of the 2021 ACM International Conference on Web Search and Data Mining*, pages 355–363.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_Eval: An Extremely Fast Python Interface to TREC_EVAL. In *Proceedings of the 2018 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 873–876.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A Survey of Preference-Based Reinforcement Learning Methods. *Journal of Machine Learning Research*, 18(136):1–46.

Zeqiu Wu, Yi Luan, Hannah Rashkin, D. Reitter, and Gaurav Singh Tomar. 2021. Conqrr: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-Training with Noisy Student Improves ImageNet Classification. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, and Mingyuan Zhou. 2024. Preference-Grounded Token-Level Guidance for Language Model Fine-Tuning. In *Proceedings of the 2024 Conference on Neural Information Processing Systems*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5985–6006.

Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2024. Ask Optimal Questions: Aligning Large Language Models with Retriever's Preference in Conversational Search. *arXiv preprint arXiv:2402.11827*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *Proceedings of the 2021 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 2020 International Conference on Learning Representations*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-Form Open-Domain Question Answering. In *Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics*, pages 6963–6975.

A Definition of Retrieval Score

We evaluate candidate query reformulations $\{\tilde{q}_t^i\}_{i=1}^M$ using pseudo reference passages $\tilde{G}_t$ based on their retrieval scores. The retrieval score $s(\tilde{q}_t^i \mid \tilde{G}_t)$ of a candidate indicates how accurately $\tilde{q}_t^i$ retrieves the passages to contain $\tilde{G}_t$. Specifically, our assessment focuses on three dimensions: (1) *coverage* (Definition 2), reflecting how *comprehensively* reference passages are retrieved within specific cutoffs; (2) *immediacy* (Definition 3), reflecting how *early* a reference passage appears in the ranking; and (3) *concordance* (Definition 4), reflecting how well the ranking *aligns* with the ideal relevance ordering of reference passages. Finally, these three scores are combined in a single weighted value (Definition 5).

Definition 2. (COVERAGE SCORE) The coverage score $s_{\text{cov}}(\tilde{q}_t^i \mid \tilde{G}_t)$ is computed as

$$s_{\text{cov}}(\tilde{q}_t^i \mid \tilde{G}_t) = \frac{1}{|K|} \sum_{k \in K} \text{Recall@k}\big(R(\tilde{q}_t^i), \tilde{G}_t\big), \quad (9)$$

where K is a predefined set of cutoff values.

Definition 3. (IMMEDIACY SCORE) The immediacy score $s_{\text{imm}}(\tilde{q}_t^i \mid \tilde{G}_t)$ is computed as

$$s_{\text{imm}}(\tilde{q}_t^i \mid \tilde{G}_t) = \text{MRR}\big(R(\tilde{q}_t^i), \tilde{G}_t\big). \quad (10)$$

Definition 4. (CONCORDANCE SCORE) The concordance score $s_{\text{con}}(\tilde{q}_t^i \mid \tilde{G}_t)$ is computed as,

$$s_{\text{con}}(\tilde{q}_t^i \mid \tilde{G}_t) = \text{NDCG}\big(R(\tilde{q}_t^i), \tilde{G}_t\big). \quad (11)$$

Definition 5. (RETRIEVAL SCORE) For each candidate $\tilde{q}_t^i$, the retrieval score $s(\tilde{q}_t^i \mid \tilde{G}_t)$ is a weighted sum of the coverage, immediacy, and concordance scores, as

$$s(\tilde{q}_t^i \mid \tilde{G}_t) = \omega_1 s_{\text{cov}}(\tilde{q}_t^i \mid \tilde{G}_t) + \omega_2 s_{\text{imm}}(\tilde{q}_t^i \mid \tilde{G}_t) + \omega_3 s_{\text{con}}(\tilde{q}_t^i \mid \tilde{G}_t), \quad (12)$$

where $\omega_1, \omega_2, \omega_3 \geq 0$ and $\omega_1 + \omega_2 + \omega_3 = 1$.

Details on the retrieval evaluation metrics are provided in Appendix C.1.

B Dataset Details

B.1 General-Domain: QReCC and TopiOCQA

The QReCC dataset (Anantha et al., 2021) contains 14K multi-turn conversations with a total of 80K question-answer pairs, aiming to retrieve reference passages from a large corpus of 54 million passages. Similarly, the TopiOCQA dataset (Adlakha et al., 2022) includes 3.9K conversations featuring topic shifts, comprising 51K question-answer pairs. Its passage collection is derived from Wikipedia and consists of approximately 20 million passages. For both datasets, small random subsets of the training data were used to construct the validation sets. While these datasets are well-suited for general-domain conversational contexts, they lack focus on domain-specific applications such as scientific question answering.

B.2 Specialized-Domain: SciConvQA

Information-seeking conversations span a wide range of domains, from general topics to specialized areas like science, reflecting diverse user interests. To evaluate existing CQR methods and DUALREFORM, we introduce the SciConvQA dataset, composed of information-seeking conversations generated from renowned scientific journals.

The conversation generation process follows the protocol described in Appendix A of the TopiOCQA (Adlakha et al., 2022) paper, which provides the methodology for creating conversational datasets. While TopiOCQA relies on crowdsourced annotations, it incurs high costs or risks of diminished quality when applied to specialized scientific domains. Hence, we employ gpt-4o-2024-08-06 (OpenAI, 2023) for automated conversation generation, followed by post-hoc manual quality validation. Overall, the conversation generation process involves two steps: (1) selecting a scientific journal as the seed topic and (2) generating questioner-answerer interactions. Table 14 shows a representative conversation from SciConvQA.

Seed Topics and Document Collection. SciConvQA is constructed using scientific journal data provided by the Korea Institute of Science and Technology Information (KISTI), a government-funded research institute. The scientific journal dataset, accessible at `https://aida.kisti.re.kr/data/b22c73ed-fa19-47b0-87b3-a509df8380e5`, includes a total of 481,578 academic articles, comprising both Korean and English publications. Detailed information about the dataset construction is available at the linked source. For our study, we utilize 120,916 English articles to construct the external database corpus, from which a subset is sampled to generate conversations. These articles span 749 diverse scientific fields, including biology, medicine, and architecture.

Conversation Generation. We modify the conversation annotation protocol of TopiOCQA to design a prompt for gpt-4-1106-preview, including an in-context demonstration to illustrate the conversation generation process based on a seed topic. The prompt template with its demonstration is shown in Figures 11–13. During the conversation generation process, each article serves as a seed topic, and the reference passages for conversation turns are selected from the article. Specifically, for each conversation turn, a reference passage ("rationale" in the prompt) is selected as a substring of the article's content that justifies the answer, recorded directly below the corresponding answer, as the demostrastive conversation in Figure 13.

Passage Database Construction. The passage database is constructed using 120,916 English articles as the retrieval target. Specifically, we employ Langchain's `RecursiveCharacterTextSplitter` with a chunk size of 500 and a chunk overlap of 100 (LangChain, 2025), resulting in a database consisting of 1,909,524 passages.

Post-Processing. To ensure compatibility with existing datasets (e.g., TopiOCQA), we standardize the format of the raw conversations generated during the conversation generation process. Due to inconsistencies in the output structure of chat completions, we extract only the relevant content using a custom post-processing pipeline. Furthermore, the passage ID for the ideal reference passage corresponding to each query is assigned by identifying the longest common substring between the generated "rationale" and passages in the external passage database. The entire implementation of the post-processing procedure is provided in the DUALREFORM's code repository.

B.3 Exploratory Analysis for SciConvQA

Data Statistics. Table 7 presents the statistics of the SciConvQA dataset. In summary, there are 11,953 turns across 900 conversations, with an average of 11.53 words per query, 15.59 words per response, and 13.28 turns per conversation.

Domain Similarity Comparison of Conversational Datasets. The t-SNE visualization in Figure 8 offers a qualitative insight into the similarity among three conversational datasets: QReCC, TopiOCQA, and SciConvQA. In general, QReCC and TopiOCQA show significant overlap in the embedding space, suggesting high semantic similarity between these datasets. This is attributed to their common focus on general-domain conversational topics. In contrast, SciConvQA forms a clearly separate cluster due to its specialized domain (e.g., scientific topics), reflecting the domain difference from the other two datasets. In real-world applications, CQR models are required to handle such diverse datasets across general-domain and specialized-domain conversations.

Dataset	Train	Test	Overall
# Turns	9,999	1,954	11,953
# Conversations	750	150	900
# Words / Query	11.51	11.62	11.53
# Words / Response	15.54	15.83	15.59
# Turns / Conversation	13.33	13.03	13.28

Table 7: Dataset statistics of SciConvQA

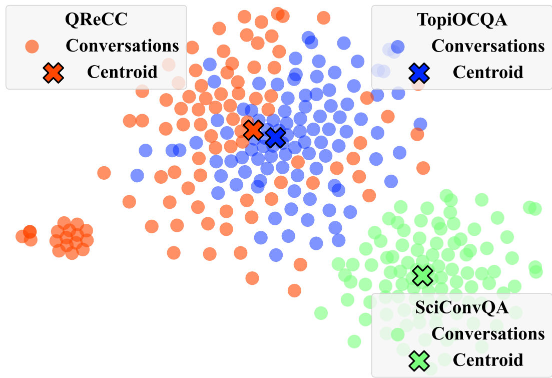


Figure 8: t-SNE visualization (Van der Maaten and Hinton, 2008) of three conversational datasets: QReCC, TopiOCQA, and SciConvQA. For each dataset, 100 randomly sampled conversations were encoded using the pretrained Sentence Transformer (Reimers and Gurevych, 2019; Ni et al., 2022; Ye et al., 2023) and projected into 2D embedding space via t-SNE. The conversations (denoted by ● symbols) from the same dataset are in the same color, where the centroid of each dataset's conversations is denoted by a × symbol.

C Experiment Details

C.1 Evaluation Metrics

Metrics for Retrieval Accuracy. We employ three widely used metrics (Qu et al., 2020; Yu et al., 2021; Lin et al., 2021; Ye et al., 2023): MRR, NDCG@3, and Recall@k. MRR evaluates how well the system ranks the *first* relevant result, with higher scores indicating that a reference passage appears earlier in the ranked list. NDCG@3 measures the overall *alignment* of ranking with with the ideal relevance ordering of reference passages by prioritizing that they are positioned closer to the top of the retrieved passage list. Recall@k assesses the *coverage*, capturing the fraction of reference passages retrieved within the top k results. Together, these metrics provide a holistic assessment of the system's ability to perform accurate and relevant passage retrieval.

Metrics for Generation Accuracy. We employ widely used metrics (Zhang et al., 2024; Baek et al., 2023; Asai et al., 2024; Mallen et al., 2023; Jeong et al., 2024; Chirkova et al., 2024): LLMeval, ROUGE, and BertScore. LLMEval employs gpt-4o-2024-08-06 (OpenAI, 2024) as the evaluator, providing human-aligned relevance and generative quality assessments. ROUGE, specifically ROUGE-1 and ROUGE-L, evaluates lexical and structural alignment through unigram overlap and longest common subsequences. BERTScore (Zhang et al., 2020; Devlin et al., 2018) uses contextual embeddings to measure semantic similarity, enabling robust evaluation beyond surface-level matching.

C.2 Implementation Details

All baseline models are trained using their official repositories (Lin et al., 2020; Gao et al., 2023; Mo et al., 2023; Ye et al., 2023; Mao et al., 2023), with the exception of RetPo, which we implement because no public code is available.

RetPo's hyperparameters are tuned via grid search for both SFT and DPO: SFT is trained for one epoch with a learning rate of 2×10^{-5} and a batch size of 32, while DPO uses $\beta = 0.1$ (chosen from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$) and trains for two epochs under the same learning rate and batch size. Our implementation of RetPo achieves improved performance over the authors' results.

We adopt the same SFT configuration for DUALREFORM but set $\beta = 0.5$ during DPO to mitigate overfitting to initial pseudo reference passages. For other hyperparameters, DUALREFORM updates pseudo reference passages every epoch, repeating this process three times following Xie et al. (2020), and selects the top-3 relevant passages per query-response turn. Hyperparameter sensitivity analyis is provided in Appendix D.7.

For backbone models, LLM-IQR, HyDE-LLM, and LLM4CS-CoT use gpt-3.5-turbo-0125 (OpenAI, 2022), while RetPo, HyDE-FT, and DUALREFORM use Llama3.1-8b-instruct (Dubey et al., 2024). T5QR and ConvGQR follow their official T5-base implementations.

All models were implemented in PyTorch 2.1.2 and trained on NVIDIA RTX A6000 Ada GPUs. Upon acceptance, the codebase and datasets, including our re-implementation of RetPo, will be made publicly available.

Details of Retrieval Systems. We use Pyserini (Johnson et al., 2019) and Faiss (Johnson et al., 2019) for BM25 (Robertson et al., 2009) and GTR (Ni et al., 2022) retrieval systems, respectively. For BM25, we adopt the parameter settings from previous studies (Mo et al., 2023; Yoon et al., 2024; Ye et al., 2023), configuring $k_1 = 0.82$, $b = 0.68$ for QReCC, and $k_1 = 0.9$, $b = 0.4$ for TopiOCQA and SciConvQA, where k_1 adjusts term frequency normalization and b controls the impact of document frequency. For GTR[4], the maximum token length is set to 384 for both the reformulated query and passage.

Both sparse and dense retrieval systems retrieve the top-100 relevant passages per query, and the aforementioned metrics are computed using *pytrec-eval* (Van Gysel and de Rijke, 2018).

Details of Response Generation. We employ Llama-3.1-8b-instruct (Dubey et al., 2024) as the response generator. The top-4 most relevant passages, retrieved using BM25, are appended to the original query. The input query to BM25 is obtained by applying various CQR methods.

Candidate Query Generation. To generate diverse candidate queries, we build on prior studies (Yoon et al., 2024; Lai et al., 2024). Specifically, we utilize the gpt-3.5-turbo-0125 (OpenAI, 2022) via the OpenAI API[5] to transform user queries in conversational datasets into diverse candidate queries. The model is configured with a temperature of 0.8 and a top-p value of 0.8 to promote diversity, with a maximum token limit set to 2560.

We adopt two prompting strategies: Question Rewriting and Query Expansion. Question Rewriting generates 12 candidate queries, whereas Query Expansion produces 3 additional candidates by applying Llama3.1-8b-instruct to the outputs of Question Rewriting. Both strategies are applied consistently across all datasets. The prompt templates for Question Rewriting (adapted from Ye et al. (2023)) and Query Expansion (adapted from Yoon et al. (2024)) are illustrated in Figure 14 and Figure 15, respectively.

[4] https://huggingface.co/sentence-transformers/gtr-t5-large

[5] https://platform.openai.com/docs/models/gpt-3-5-turbo

Target Dataset	Query Reformulations	Sparse Retriever MRR	NDCG	R@5	R@20	Dense Retriever MRR	NDCG	R@5	R@20
QReCC	Upper Bound	50.32	46.94	58.97	78.96	51.63	49.11	63.13	79.64
	LLM-IQR	41.82	38.88	52.58	71.95	48.09	45.25	**62.13**	80.24
	HyDE-LLM	42.26	39.21	52.93	72.44	48.20	45.46	61.97	79.85
	LLM4CS-CoT	47.51	44.25	56.64	78.96	**48.53**	45.49	58.54	79.64
	T5QR	32.19	29.17	40.18	61.94	41.21	38.18	54.63	73.45
	SciConvQA ↓ *QReCC* ConvGQR	32.49	29.66	41.36	59.47	36.86	34.16	48.94	67.72
	HyDE-FT	38.87	36.09	47.41	64.62	41.23	37.92	52.55	68.76
	RetPo	39.22	36.49	48.76	65.07	45.44	42.73	59.51	78.32
	DUALREFORM	**48.40**	**45.30**	**58.91**	**77.90**	47.58	**46.98**	59.60	**80.33**

Table 8: Retrieval performance comparison of DUALREFORM against representative CQR baselines on the target dataset, QReCC. The best results (excluding Upper Bound) are highlighted in bold.

Data	Refine Methods	Pseudo Ref. Acc. NDCG	R@20	Retrieval Acc. NDCG	R@20
SciConvQA	Llama	35.62	59.37	16.53	34.03
	Llama+ICL	43.99	66.00	17.56	42.32
	DUALREFORM	**49.36**	**71.27**	**18.88**	**42.78**
TopiOCQA	Llama	35.42	50.40	22.98	53.62
	Llama+ICL	50.24	67.86	25.76	54.02
	DUALREFORM	**56.23**	**76.82**	**26.57**	**59.03**

Table 9: Comparison of response refinement methods, evaluated using the sparse retriever. The highest values are emphasized in bold.

Data	Backbone Models	Refine Methods	Pseudo Reference Accuracy MRR	NDCG	R@5	R@20
SciConvQA	**Llama3.1-70B-inst**	Llama	47.61	46.91	57.14	68.28
		Llama+ICL	49.64	48.87	58.46	69.90
	Llama3.1-8B-inst	**DUALREFORM**	**50.05**	**49.36**	**59.75**	**71.27**
TopiOCQA	**Llama3.1-70B-inst**	Llama	51.00	50.51	60.74	71.09
		Llama+ICL	55.38	55.10	64.70	73.74
	Llama3.1-8B-inst	**DUALREFORM**	**56.50**	**56.23**	**66.79**	**76.82**

Table 10: Comparison with the substantially larger Llama3.1-70b-inst backbone. Both Llama and Llama+ICL use Llama3.1-70b-inst, while DUALREFORM employs the smaller Llama3.1-8b-inst for CQR. The highest values are emphasized in bold.

D Complete Experiment Results

D.1 Extended Results: QReCC

Table 8 extends the results of Table 2 by additionally using the QReCC dataset as the target dataset. The results further highlight the importance of reference-free preference optimization by demonstrating that Upper Bound and RetPo, which leverage preference optimization, exhibit divergent performance depending on the availability of reference passages from the target dataset. Notably, DUALREFORM establishes itself as an effective approach for *reference-free* preference optimization, outperforming the baseline methods and achieving performance levels comparable to Upper Bound.

D.2 Extended Results: Effect of Pseudo Reference Refinement via CQR

Table 9 extends the results of Table 3 by additionally using other evaluation metrics, NDCG@3 and Recall@20, in order to offer a more comprehensive assessment of retrieval performance.

Comparison with the Substantially Larger Backbone Language Model, Llama3.1-70b-inst. In the Llama and Llama+ICL variants, we replace their backbones with a larger backbone, *Llama3.1-70b-inst*, whereas DUALREFORM keeps using *Llama3.1-8b-inst* as the backbone for its CQR model. As shown in Table 10, DUALREFORM achieves comparable or marginally superior performance despite using a substantially smaller language model as its backbone. This finding demonstrates that CQR can effectively substitute widely adopted LLMs for response refinement without introducing additional computational overhead.

Demonstrations for Llama+ICL. *Llama+ICL* differs from *Llama* by utilizing in-context demonstrations, as shown in Figure 16.

D.3 Extended Results: Effect of Iterative Refinement for Pseudo Reference

Table 11 extends the results of Table 4 by additionally using other evaluation metrics, NDCG@3 and Recall@20. These results again demonstrate that DUALREFORM benefits from the iterative refinement process of pseudo reference passages.

Data	Pseudo Ref. Updates	Pseudo Ref. Acc. NDCG	Pseudo Ref. Acc. R@20	Retrieval Acc. NDCG	Retrieval Acc. R@20
SciConvQA	1	37.48	58.98	15.50	40.23
	2	45.42	65.88	17.96	40.38
	3	**49.36**	**71.27**	**18.88**	**42.78**
TopiOCQA	1	38.82	50.75	23.66	52.11
	2	54.86	75.28	26.42	57.25
	3	**56.23**	**76.82**	**26.57**	**59.03**

Table 11: Effect of iterative optimization within DUALREFORM, evaluated using the sparse retriever. The highest values are emphasized in bold.

Variants	SciConvQA NDCG	SciConvQA R@20	TopiOCQA NDCG	TopiOCQA R@20	*Degrade*
No Template	44.56	65.70	54.23	73.45	*6.88%*
No Response	46.59	68.72	49.93	71.71	*7.35%*
No QE	42.84	62.75	53.94	72.57	*9.72%*
No QR	45.90	66.33	46.58	66.58	*12.77%*
DUALREFORM	**49.36**	**71.27**	**56.23**	**76.82**	-

Table 12: Effect of the query-forming template on pseudo reference accuracy using the sparse retriever. The highest values are emphasized in bold.

Context: [Q: who plays the general in the incredible hulk A: Thaddeus "Thunderbolt" Q: when was the movie released? A: June 13, 2008 Q: what is the storyline? A: Bruce Banner becomes the Hulk as an unwitting pawn in a military scheme to reinvigorate the "Super-Soldier" program through gamma radiation. Banner goes on the run from the military while attempting to cure himself of the Hulk. Q: who are the other characters ? A: General Thaddeus "Thunderbolt" Ross, Betty, Dr. Samuel Sterns Q: who portrayed the protagonist of the movie? A: Bruce Banner Q: where did the filming take place? A: Toronto Q: who did the soundtrack ? A: Craig Armstrong Q: which production company is associated with it? A: Marvel Studios Q: can you name any other movies they produced?] Question: Can you clearly state the main points of the last answer ('Iron Man, X-Men, Spider-Man'), contextualizing them and resolving coreferences?

Figure 9: Top attention weights assigned by DUALREFORM during response refinement for the conversation in Figure 7, with high-weight regions highlighted in red.

D.4 Effect of Query-Forming Template

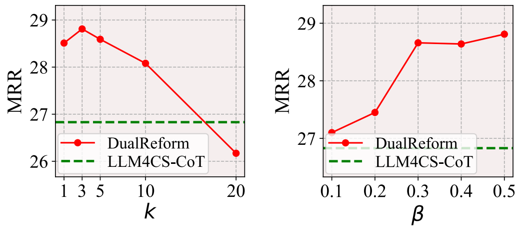
Table 12 extends the results of Table 5 by additionally using other evaluation metrics, NDCG@3 and Recall@20. These results further validate that the performance of the variants consistently declines compared to DUALREFORM, confirming the effectiveness of its query-forming template.

D.5 More Examples of Refined Responses

Figure 17 complements Figure 7, presenting the complete refined responses generated by various methods. Figure 9 illustrates the top attention weights assigned by DUALREFORM during response refinement, demonstrating its ability to correctly attend to the relevant context (e.g., “Marvel Studios”) and the target response (e.g., “Iron Man, X-Men, Spider-Man”). More examples for other conversations are provided in Figures 18–20.

CQR Methods	Generation Accuracy LLMEval	ROUGE-1	ROUGE-L	BertScore
LLM-IQR	25.34	20.36	19.09	84.32
HyDE-LLM	27.36	24.83	23.30	85.28
LLM4CS-CoT	33.78	29.20	27.78	85.88
T5QR	18.92	17.07	16.50	83.67
ConvGQR	21.62	22.02	20.89	84.84
HyDE-FT	15.98	15.37	14.72	84.01
RetPo	33.11	26.01	24.39	85.74
DUALREFORM	**35.81**	**31.63**	**30.10**	**86.48**

Table 13: Response generation accuracy with passages retrieved via different CQR methods on TopiOCQA. The highest values are emphasized in bold.



(a) # of Pseudo Ref. (b) DPO Regularization.

Figure 10: Effects of the hyperparameters of DUALREFORM on MRR. The green dashed line denotes the performance of LLM4CS-CoT, the strongest baseline.

D.6 Generation Accuracy on TopiOCQA

Table 13 extends the results of Table 6 by additionally using the TopiOCQA dataset. These results show that DUALREFORM consistently outperforms the baselines by achieving higher generation accuracy through accurate passage retrieval across diverse conversational domains.

D.7 Parameter Sensitivity Analysis

We conduct the sensitivity analysis of DUALREFORM's hyperparameters, specifically the number of pseudo reference passages k in Definition 1 and the regularization parameter β of DPO in Eq. (8).

Number of Pseudo Reference Passages k. Figure 10(a) presents the impact of k on retrieval performance. The parameter k controls the number of top-ranked passages used as pseudo reference passages. Lower values selectively include only the most relevant passages, while higher values introduce additional but less relevant passages. In general, the retrieval accuracy stabilizes between 1 and 5, after which it declines, indicating a negative impact from less relevant passages beyond a specific threshold.

DPO Regularization Parameter β. Figure 10(b) presents retrieval accuracy across different values of $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, as guided by prior works (Rashul et al., 2023; Furuta et al., 2024).

This parameter controls the trade-off between aligning the model with user preferences and retaining the behavior of the pre-trained model. Lower values prioritize the former, while higher values place greater emphasis on the latter. In general, increasing β improves performance, with a plateau observed around 0.3 and a peak at 0.5. This trend is attributed to the ability of higher β values to reduce overfitting to the initial pseudo reference passages, further enabling model refinements in DUALREFORM through iterative optimization.

SciConvQA conversation
Q: How do genomic tools enhance animal breeding programs?
A: Genomic tools, such as single nucleotide polymorphism (SNP), have led to a new method known as "genomic selection." This method utilizes dense SNP genotypes covering the entire genome to predict the breeding value.
Q: Which application is used in evaluating these programs?
A: ZPLAN+ is used to evaluate and optimize these programs.
Q: What parameters does it consider?
A: It considers genetic and economic parameters.
Q: Can you tell me more about the types of strategies it models?
A: It models four selection strategies: the current conventional program and three based on genomic information.
Q: What's unique about the final approach among them?
A: The final approach, GS3, is unique because it combines pedigree, genomic enhanced breeding values, performance, and progeny information.
Q: How were the male candidates evaluated in this scheme?
A: They were genotyped and their selection was based on performance tests and GEBV.
Q: Do we have information about the costs associated with these methods?
A: Yes, the cost of genotyping was assumed to be $120 per pig.
Q: How does this cost compare to more traditional testing methods?
A: Performance testing costs $55 per tested pig, which is less than genotyping.
Q: How many candidates undergo this evaluation when adventurers start the process?
A: Initially, 1,000 male candidates were considered in the genomic selection process.
Q: In the final phase of selection, how many of these males are retained?
A: In the final phase, 23 senior boars were retained.
Q: How does the conventional program handle progeny information differently than genomics?
A: The conventional program uses progeny records without considering genetic marker information.
Q: Can you summarize which traits measured the field performance?
A: Traits measured include average daily gain, back fat thickness, and feed conversion rate.
Q: Are these the same for testing the station?
A: No, station tests focus on meat quality traits like pH, meat color (L*), and intra-muscular fat.
Q: In this optimized workflow, what benefit is sought above all?
A: The primary goal is high genetic gains with low breeding costs.

Table 14: Example of a SciConvQA conversation. The conversation is generated from Lopez et al. (2016), published in a renowned journal and stored as 'JAKO201614137726690.json' in the scientific journal dataset described in Section B.2.

Create an information-seeking conversation between two annotators: a questioner and an answerer. We give you multiple sections of an academic paper as seed topics for the information-seeking questions.
Assume that the questioner has access only to the main topic of the given content, while the answerer can access the full text. Allow topic switching by enabling the answerer to refer to sections from different sections of papers.

The annotators are provided with the following guidelines.
Guidelines for the questioner:

- The first question should be unambiguous and clear about the main topic of academic paper. The questioner, only knowing the main topic, cannot ask directly about the study's focus. Thus, do not ask like "What is the primary focus of the study?" as 1st question.
- The follow-up questions are contextualized and always dependent on the conversation history (especially, last answerer's respond) so that question itself is hard to understand.
- Avoid using same words as in section titles of the document. E.g. if the section title is "Awards", a plausible question can be "What accolades did she receive for her work?".
- The conversation should involve multiple documents (topics).

Guidelines for the answerer:

- Based on the question, identify the relevant document and section.
- The answer should be based on the contents of the identified document.
- The rationale should be a sub-string of content such that it justifies the answer and should be recorded below the answers.
- The answer should be a sub-string in rationale whenever possible. However, answers should be edited to fit the conversational context (adding yes, no), perform reasoning (e.g. counting) etc.
- Personal opinions should never be included.

- **Example:**
  - **Content:** {content}
  - **Information-seeking conversation:** {conversation}

- **An information-seeking conversation:**
  - **Content:** {topic seed}

Write an information-seeking conversation comprising 10-15 QA turns based on the provided content. The questions should employ co-references rather than explicit identifiers (e.g., names, titles, locations) to ensure contextual dependence on prior turns. Do not replicate content from demonstrative examples.

Figure 11: Prompt used for SciConvQA dataset generation. The example of {content} is provided in Figure 12, and the example of {conversation} can be found in Figure 13.

The Byzantine Empire, also referred to as the Eastern Roman Empire, was the continuation of the Roman Empire centered in Constantinople during Late Antiquity and the Middle Ages. The eastern half of the Empire survived the conditions that caused the fall of the West in the 5th century AD, and continued to exist until the fall of Constantinople to the Ottoman Empire in 1453. During most of its existence, the empire remained the most powerful economic, cultural, and military force in the Mediterranean world. The term "Byzantine Empire" was only coined following the empire's demise; its citizens referred to the polity as the "Roman Empire" and to themselves as "Romans". Due to the imperial seat's move from Rome to Byzantium, the adoption of state Christianity, and the predominance of Greek instead of Latin, modern historians continue to make a distinction between the earlier Roman Empire and the later Byzantine Empire.
The empire was largely dismantled in 1204, following the Sack of Constantinople by Latin armies at the end of the Fourth Crusade; its former territories were then divided into competing Greek rump states and Latin realms. Despite the eventual recovery of Constantinople in 1261, the reconstituted empire would wield only regional power during its final two centuries of existence. Its remaining territories were progressively annexed by the Ottomans in perennial wars fought throughout the 14th and 15th centuries. The fall of Constantinople to the Ottomans in 1453 ultimately brought the empire to an end. Many refugees who had fled the city after its capture settled in Italy and throughout Europe, helping to ignite the Renaissance. The fall of Constantinople is sometimes used to mark the dividing line between the Middle Ages and the early modern period.
The situation became worse for Byzantium during the civil wars after Andronikos III died. A six-year-long civil war devastated the empire, allowing the Serbian ruler Stefan Dušan to overrun most of the empire's remaining territory and establish a Serbian Empire. In 1354, an earthquake at Gallipoli devastated the fort, allowing the Ottomans (who were hired as mercenaries during the civil war by John VI Kantakouzenos) to establish themselves in Europe. By the time the Byzantine civil wars had ended, the Ottomans had defeated the Serbians and subjugated them as vassals. Following the Battle of Kosovo, much of the Balkans became dominated by the Ottomans.
Constantinople by this stage was underpopulated and dilapidated. The population of the city had collapsed so severely that it was now little more than a cluster of villages separated by fields. On 2 April 1453, Sultan Mehmed's army of 80,000 men and large numbers of irregulars laid siege to the city. Despite a desperate last-ditch defence of the city by the massively outnumbered Christian forces (c. 7,000 men, 2,000 of whom were foreign), Constantinople finally fell to the Ottomans after a two-month siege on 29 May 1453. The final Byzantine emperor, Constantine XI Palaiologos, was last seen casting off his imperial regalia and throwing himself into hand-to-hand combat after the walls of the city were taken.
Mehmed continued his conquests in Anatolia with its reunification and in Southeast Europe as far west as Bosnia. At home, he made many political and social reforms. He encouraged the arts and sciences, and by the end of his reign, his rebuilding program had changed Constantinople into a thriving imperial capital. He is considered a hero in modern-day Turkey and parts of the wider Muslim world. Among other things, Istanbul's Fatih district, Fatih Sultan Mehmet Bridge and Fatih Mosque are named after him.
Anatolia (Turkish: Anadolu), also known as Asia Minor, is a large peninsula or a region in Turkey, constituting most of its contemporary territory. Geographically, the Anatolian region is bounded by the Mediterranean Sea to the south, the Aegean Sea to the west, the Turkish Straits to the north-west, and the Black Sea to the north. The eastern and southeastern boundary is either the southeastern and eastern borders of Turkey, or an imprecise line from the Black Sea to Gulf of Iskenderun. Topographically, the Sea of Marmara connects the Black Sea with the Aegean Sea through the Bosporus strait and the Dardanelles strait, and separates Anatolia from Thrace in the Balkan peninsula of Southeastern Europe.
The Akkadian Empire (/əˈkeɪdiən/) was the first known ancient empire of Mesopotamia, succeeding the long-lived civilization of Sumer. Centered on the city of Akkad (/ˈækæd/) and its surrounding region, the empire united Akkadian and Sumerian speakers under one rule and exercised significant influence across Mesopotamia, the Levant, and Anatolia, sending military expeditions as far south as Dilmun and Magan (modern United Arab Emirates, Saudi Arabia, Bahrain, Qatar and Oman) in the Arabian Peninsula.

Figure 12: Example of {content} in Figure 11.

```
Q1: when was the byzantine empire born what was it originally called?
A1: 5th century AD and was called Eastern Roman Empire, or Byzantium
rationale: The Byzantine Empire, also referred to as the Eastern Roman Empire, was the continuation of the Roman Empire centered in Constantinople during Late Antiquity and the Middle Ages. The eastern half of the Empire survived the conditions that caused the fall of the West in the 5th century AD, and continued to exist until the fall of Constantinople to the Ottoman Empire in 1453.

Q2: and when did it fall?
A2: 1453
rationale: The fall of Constantinople to the Ottomans in 1453 ultimately brought the empire to an end.

Q3: which battle or event marked the fall of this empire?
A3: A six-year-long civil war followed by attack from Sultan Mehmed’s army
rationale: A six-year-long civil war devastated the empire;  On 2 April 1453, Sultan Mehmed's army of 80,000 men and large numbers of irregulars laid siege to the city.

Q4: did he conquer other territories as well?
A4: Yes. Anatolia and in Southeast Europe as far west as Bosnia
rationale: Mehmed continued his conquests in Anatolia with its reunification and in Southeast Europe as far west as Bosnia

Q5: where is the first area located in present day terms?
A5: Turkey
rationale: Anatolia (Turkish: Anadolu), also known as Asia Minor,[a] is a large peninsula or a region in Turkey

Q6: who were the oldest known inhabitants of this region?
A6: Mesopotamian-based Akkadian Empire
rationale: The Akkadian Empire (/əˈkeɪdiən/)[2] was the first known ancient empire of Mesopotamia
```

Figure 13: Examples of {conversation} in Figure 11.

Given a question, its previous questions (Q) and answers (A), decontextualize the question by addressing coreference and omission issues. The resulting question should retain its original meaning and be as informative as possible, and should not duplicate any previously asked questions in the context.

Context: [Q: When was Born to Fly released? A: Sara Evans's third studio album, Born to Fly, was released on October 10, 2000.]
Question: Was Born to Fly well received by critics?
Rewrite: Was Born to Fly well received by critics?

Context: [Q: When was Keith Carradine born? A: Keith Ian Carradine was born August 8, 1949. Q: Is he married? A: Keith Carradine married Sandra Will on February 6, 1982.]
Question: Do they have any children?
Rewrite: Do Keith Carradine and Sandra Will have any children?

Context: [Q: Who proposed that atoms are the basic units of matter? A: John Dalton proposed that each chemical element is composed of atoms of a single, unique type, and they can combine to form more complex structures called chemical compounds.]
Question: How did the proposal come about?
Rewrite: How did John Dalton's proposal that each chemical element is composed of atoms of a single unique type, and they can combine to form more complex structures called chemical compounds come about?

Context: [Q: What is it called when two liquids separate? A: Decantation is a process for the separation of mixtures of immiscible liquids or of a liquid and a solid mixture such as a suspension. Q: How does the separation occur? A: The layer closer to the top of the container-the less dense of the two liquids, or the liquid from which the precipitate or sediment has settled out-is poured off.]
Question: Then what happens?
Rewrite: Then what happens after the layer closer to the top of the container is poured off with decantation?

Context: {context}
Question: {query}
Rewrite:

Figure 14: Prompt used for Query Rewriting.

```
Given a question, please answer the question in a sentence. The answer should be as informative as possible.

Question: and by whom was the game the last of us established?
Answer: Andy Gavin and Jason Rubin. Naughty Dog, LLC (formerly JAM Software, Inc.) is an American first-party
video game developer based in Santa Monica, California. Founded by Andy Gavin and Jason Rubin in 1984 as an
independent developer.

Question: is chelsea a club?
Answer: Yes, chelsea is an English professional football club.

Question: is call me by your name a movie?
Answer: Yes, based on a book of the same name. Call Me by Your Name is a 2017 coming-of- age romantic drama
film directed by Luca Guadagnino. Its screenplay, by James Ivory, who also co-produced, is based on the 2007
novel of the same name by Andr Aciman.

Question: where was ulysses s. grant from?
Answer: Hiram Ulysses Grant was born in Point Pleasant, Ohio, on April 27, 1822, to Jesse Root Grant, a tanner
and merchant, and Hannah Simpson Grant.

Question: {rewritten query}
Answer:
```

Figure 15: Prompt used for Query Expansion.

```
Context: [Q: When was Born to Fly released? A: Sara Evans's third studio album, Born to Fly, was released on
October 10, 2000.]
Explain: The last answer provides the release date of Sara Evans's third studio album, 'Born to Fly,' which
was October 10, 2000.

Context: [Q: When was Keith Carradine born? A: Keith Ian Carradine was born August 8, 1949. Q: Is he married?
A: Keith Carradine married Sandra Will on February 6, 1982.]
Explain: The last answer indicates that Keith Carradine married Sandra Will on February 6, 1982. In the
context of the overall conversation, the focus shifts from Keith Carradine's birth date to his marital status,
specifically addressing whether he is married.

Context: [Q: I've been curious about studies improving livestock productivity; is there a specific goal that
they often focus on? A: Undoubtedly, enhancing the efficiency of livestock breeds while minimizing their
ecological footprint are important goals in animal farming. Q: You mentioned improving certain efficiencies.
Which one takes up most of the cost in this setting? A: In animal agriculture, about 70% of the total
production expenses are attributed to feed costs.]
Explain: The last answer in this conversation states that in animal agriculture, about 70% of the total
production expenses are attributed to feed costs. This means that the majority of the expenses incurred in
livestock production are related to providing feed for the animals.
```

Figure 16: Demonstrations used for the Llama+ICL variant.

<table>
<tr><th>Method</th><th>Refined Responses $\tilde{a}_9$</th><th>Conversation</th></tr>
<tr><td>Reference Passage</td><td>Marvel Studios Introduction Marvel Studios, LLC (originally known as Marvel Films from 1993 to 1996) is an American film and television studio that is a subsidiary of The Walt Disney Studios. Marvel Studios is known for the production of the Marvel Cinematic Universe films, based on characters that appear in Marvel Comics publications. Marvel Studios has released 23 films since 2008 within the Marvel Cinematic Universe, from "Iron Man" (2008) to "Spider-Man: Far From Home" (2019). These films all share continuity with each other, along with the One-Shots produced by the studio and the television shows produced by Marvel Television. In addition to the Marvel Cinematic Universe, Marvel Studios was also involved with the production of other Marvel-character film franchises that have exceeded $1 billion in North American box office revenue: the X-Men and Spider-Man multi-film franchises.</td><td rowspan="5">< Conv. History >
q_1: who plays the general in the incredible hulk?
a_1: Thaddeus "Thunderbolt".
q_2: when was the movie released?
a_2: June 13, 2008.
q_3: what is the storyline?
a_3: Bruce Banner becomes the Hulk as an unwitting pawn in a military scheme to reinvigorate the "Super-Soldier" program through gamma radiation. Banner goes on the run from the military while attempting to cure himself of the Hulk.
q_4: who are the other characters?
a_4: General Thaddeus "Thunderbolt" Ross, Betty, Dr. Samuel Sterns.
q_5: who portrayed the protagonist of the movie?
a_5: Bruce Banner.
q_6: where did the filming take place?
a_6: Toronto.
q_7: who did the soundtrack?
a_7: Craig Armstrong
q_8: which production company is associated with it?
a_8: Marvel Studios
q_9: can you name any other movies they produced?

< Target Response >
a_9: Iron Man, X-Men, Spider-Man</td></tr>
<tr><td>Llama-ICL</td><td>The last answer in this conversation lists several production-related details about the movie "The Incredible Hulk". It names the production company associated with the movie as Marvel Studios and mentions a few other movies produced by the same company, which are Iron Man, X-Men, and Spider-Man.</td></tr>
<tr><td>No Template</td><td>Can you name other movies produced by Marvel Studios? Can you name other movies produced by Marvel Studios? … (degenerated)</td></tr>
<tr><td>No Response</td><td>Can you clearly state the main points of the last answer, contextualizing them and resolving coreferences? The main points of the last answer are: Marvel Studios produced the movie The Incredible Hulk, which was released on June 13, 2008, and had a storyline where Bruce Banner becomes the Hulk as an unwitting pawn in a military scheme to reinvigorate the "Super-Soldier" program through gamma radiation. He goes on the run from the military while attempting to cure himself of the Hulk. The movie was filmed in Toronto, and the soundtrack was composed by Craig Armstrong. Marvel Studios is associated with the production of the movie and has also produced other notable films like Iron Man, X-Men, and Spider-Man.</td></tr>
<tr><td>DualReform</td><td>Can you clearly state the main points of the movies produced by Marvel Studios, contextualizing them and resolving coreferences? The main points of the movies produced by Marvel Studios include Iron Man, X-Men, and Spider-Man, which are three popular superhero films that have been well-received by audiences and critics alike. Iron Man is a film about a billionaire inventor who becomes a superhero, X-Men is a film about a group of mutants with unique powers, and Spider-Man is a film about a young man who gains powers after being bitten by a spider. All three films have been successful and have contributed to the Marvel Cinematic Universe.</td></tr>
</table>

Figure 17: Examples of refined responses generated by different methods on TopiOCQA. Fragments strongly aligned with the reference passage are highlighted in blue, while fragments with weaker connections (e.g., off-topic elements referring to previous conversation topics) are marked in red.

<table>
<tr><th>Method</th><th>Refined Responses $\tilde{a}_{10}$</th><th>Conversation</th></tr>
<tr><td>Reference Passage</td><td>Atari, Inc. was an American video game developer and home computer company founded in 1972 by Nolan Bushnell and Ted Dabney. Primarily responsible for the formation of the video arcade and modern video game industry, the company assets were split in 1984 as a direct result of the video game crash of 1983, and the company was renamed first Atari Games, Inc. and then Atari Holdings, Inc. After 1985, when its coin-operated games division was sold to Namco and its arcade operations division was shut down, it remained a non-operating subsidiary of Warner Communications and its successor company, Time Warner, until it was merged into the parent company in 1992.</td><td rowspan="5">< Conversation History >
q_1: when did the apple iphone 4s come out?
a_1: October 14, 2011.
q_2: which company manufactures this smartphone?
a_2: Apple Inc.
q_3: anything significant about the software used in this phone?
a_3: It debuted with iOS 5, the fifth major version of iOS, Apple's mobile operating system, that introduced features including iCloud, iMessage, Notification Center, Reminders, and Twitter integration.
q_4: was the product pulled into any controversy?
a_4: Yes, on December 22, 2015, Apple faced a class action lawsuit for crippling the iPhone 4S with the iOS 9 update with slow and buggy software or to pay hundreds of dollars on a new iPhone.
q_5: explain the notification feature mentioned above?
a_5: Notification Center is a feature in iOS and macOS that provides an overview of alerts from applications, it displays notifications until the user completes an associated action, rather than requiring instant resolution.
q_6: who owns the company that manufactures the phone?
a_6: UNANSWERABLE
q_7: who were the founders?
a_7: Steve Jobs, Steve Wozniak, and Ronald Wayne.
q_8: which university did the first person attend?
a_8: Jobs enrolled at Reed College
q_9: where did he have his first job?
a_9: Atari, Inc.
q_{10}: when was this founded?

< Target Response >
a_{10}: 1972.</td></tr>
<tr><td>Llama+ICL</td><td>The last answer in this conversation states that the feature mentioned is Notification Center, which is a feature in iOS and macOS that provides an overview of alerts from applications, displaying notifications until the user completes an associated action, rather than requiring instant resolution.</td></tr>
<tr><td>No Template</td><td>When was Atari, Inc. founded in 1972? Atari, Inc. was founded in 1972.</td></tr>
<tr><td>No Response</td><td>Can you clearly state the main points of the last answer, contextualizing them and resolving coreferences? The main points of the last answer are: Steve Jobs, Steve Wozniak, and Ronald Wayne were the founders of Apple Inc., the company that manufactures the iPhone 4S. Steve Jobs enrolled at Reed College, and he had his first job at Atari, Inc., which was founded in 1972. The iPhone 4S was released on October 14, 2011, and it debuted with iOS 5, the fifth major version of iOS, which introduced features including iCloud, iMessage, Notification Center, Reminders, and Twitter integration. Apple faced a class action lawsuit in December 2015 for crippling the iPhone 4S with slow and buggy software. Notification Center is a feature in iOS and macOS that provides an overview of alerts from applications, displaying notifications until the user completes an associated action.</td></tr>
<tr><td>DualReform</td><td>Can you clearly state the main points of the founding of Atari, Inc. in 1972, contextualizing them and resolving coreferences? Atari, Inc. was founded in 1972 by Nolan Bushnell and Ted Dabney, and it is widely recognized as a pioneering company in the video game industry. The company was initially called Atari, Inc. and was later renamed to Atari Corporation in 1984. Atari, Inc. is known for its iconic games such as Pong, Asteroids, and Centipede, and it played a significant role in popularizing video games in the 1970s and 1980s.</td></tr>
</table>

Figure 18: Examples of refined responses generated by different methods on TopiOCQA. Fragments strongly aligned with the reference passage are highlighted in blue, while fragments with weaker connections (e.g., off-topic elements referring to previous conversation topics) are marked in red.

<table>
<tr><th>Method</th><th>Refined Responses $\tilde{a}_{10}$</th><th>Conversation</th></tr>
<tr><td>Reference Passage</td><td>The underlying bedrock or limestone of the Everglades basin affects the "hydroperiod", or how long an area within the region stays flooded throughout the year. Longer hydroperiods are possible in areas that were submerged beneath seawater for longer periods of time, while the geology of Florida was forming. More water is held within the porous ooids and limestone than older types of rock that spent more time above sea level. A hydroperiod of ten months or more fosters growth of sawgrass, whereas a shorter hydroperiod of six months or less promotes beds of periphyton, a growth of algae and other microscopic organisms.</td><td rowspan="5">< Conversation History >
q_1: what is the longest river in the everglades?
a_1: UNANSWERABLE
q_2: in which state of the us is this region located?
a_2: Florida.
q_3: what is the type of ecosystem it has in this mentioned region of this state?
a_3: It is a natural region of tropical wetlands, comprising the southern half of a large drainage basin within the Neotropical realm.
q_4: name some plants typically found in this place.
a_4: Sawgrass prairies, bladderwort, waterlily etc.
q_5: is there a botanical name of the second one?
a_5: Yes, "Utricularia"
q_6: is this a carnivorous plant?
a_6: Yes, it captures small organisms by means of bladder-like traps.
q_7: what do they feed on?
a_7: Protozoa and rotifers swimming in water-saturated soil.
q_8: what is the peculiar feature of their respiration?
a_8: They have significantly greater respiration rates than most vegetative tissue, primarily due to their complex energy-dependent traps.
q_9: how does reproduction take place in these plants?
a_9: Certain plants in particular seasons might produce closed, self-pollinating ("cleistogamous") flowers; but the same plant or species might produce open, insect-pollinated flowers elsewhere or at a different time of year, and with no obvious pattern.
q_{10}: what rocks are found in that us state in the region where these plants are found?

< Target Response >
a_{10}: Bedrock or limestone.</td></tr>
<tr><td>Llama+ICL</td><td>The last answer indicates that the rocks found in the region where the mentioned plants are found in the state of Florida are bedrock or limestone. This answer provides a piece of information about the geological characteristics of the area.</td></tr>
<tr><td>No Template</td><td>Bedrock or limestone rocks are found in the region where these plants are found in the state of Florida.</td></tr>
<tr><td>No Response</td><td>Can you clearly state the main points of the fact that certain plants in the Everglades region of Florida have self-pollinating flowers and may also produce open, insect-pollinated flowers elsewhere or at a different time of year, and that the region is characterized by the presence of bedrock or limestone rocks? The Everglades region in Florida is characterized by the presence of plants that have self-pollinating flowers and may also produce open, insect-pollinated flowers elsewhere or at a different time of year, and the region is also marked by the presence of bedrock or limestone rocks. The plants in this region, such as sawgrass prairies, bladderwort, and waterlily, have unique features like significantly greater respiration rates and complex energy-dependent traps, and certain species may produce closed, self-pollinating flowers in particular seasons, while others may produce open, insect-pollinated flowers elsewhere or at a different time of year. The Everglades region is a natural region of tropical wetlands, comprising the southern half of a large drainage basin within the Neotropical realm, and is characterized by the presence of bedrock or limestone rocks, which are found in the region where these plants are found.</td></tr>
<tr><td>DualReform</td><td>Can you clearly state the main points of the answer 'Bedrock or limestone' in the context of the Everglades region in Florida, and resolve any coreferences or ambiguities? The main points of the answer 'Bedrock or limestone' in the context of the Everglades region in Florida are that the region is characterized by bedrock or limestone, which are the primary geological formations found in the area. These formations provide a foundation for the unique ecosystem of the Everglades, supporting the growth of plants such as sawgrass, bladderwort, and waterlily, and creating a habitat for a diverse range of wildlife.</td></tr>
</table>

Figure 19: Examples of refined responses generated by different methods on TopiOCQA. Fragments strongly aligned with the reference passage are highlighted in blue, while fragments with weaker connections (e.g., off-topic elements referring to previous conversation topics) are marked in red.

<table>
<tr><th>Method</th><th>Refined Responses $\tilde{a}_9$</th><th>Conversation</th></tr>
<tr><td>Reference Passage</td><td>"A Star Is Born", Enigma, and "Chromatica" In March 2018, Gaga supported the March for Our Lives gun-control rally in Washington, D.C., and released a cover of Elton John's "Your Song" for his tribute album "Revamp". Later that year, she starred as struggling singer Ally in Bradley Cooper's musical romantic drama "A Star Is Born", a remake of the 1937 film of the same name. The film follows Ally's relationship with singer Jackson Maine (played by Cooper), which becomes strained after her career begins to overshadow his. Cooper approached Gaga after seeing her perform at a cancer research fundraiser; an admirer of Cooper's work, Gaga agreed to the project due to its portrayal of addiction and depression.</td><td rowspan="5">< Conversation History >
q_1: who is lady gaga?
a_1: Stefani Joanne Angelin. Germanotta, known professionally as Lady Gaga, is an American singer, songwriter and actress.
q_2: where was she born?
a_2: Manhattan, New York City.
q_3: when?
a_3: March 28, 1986.
q_4: can you please name some of her songs?
a_4: Bad Romance", "Telephone", and "Alejandro".
q_5: describe her music style?
a_5: Electronic rock and techno.
q_6: describe her music style?
a_6: Electronic rock and techno.
q_7: can you please name some of her albums?
a_7: "The Fame", "Born This Way", "Artpop".
q_8: has she won any awards for her music?
a_8: Yes, 11 Grammy Awards, an Academy Award, a BAFTA Award, three Brit Awards, two Golden Globe Awards, eighteen MTV Video Music Awards, various Guinness World Record.
q_9: has she acted in movies?

< Target Response >
a_9: Yes, "A Star Is Born".</td></tr>
<tr><td>Llama+ICL</td><td>In the context of the overall conversation, the last answer mentions Lady Gaga's acting career, specifically that she has acted in the movie "A Star Is Born".</td></tr>
<tr><td>No Template</td><td>Has Lady Gaga acted in movies, including "A Star Is Born"? Yes, Lady Gaga has acted in movies, including "A Star Is Born".</td></tr>
<tr><td>No Response</td><td>Can you clearly state the main points of Lady Gaga's awards and acting career? Lady Gaga has won 11 Grammy Awards, an Academy Award, a BAFTA Award, three Brit Awards, two Golden Globe Awards, eighteen MTV Video Music Awards, and various Guinness World Records for her music. Additionally, she has acted in the movie "A Star Is Born".</td></tr>
<tr><td>DualReform</td><td>Can you clearly state the main points of the answer "Yes, "A Star Is Born" in relation to Lady Gaga's acting career? Lady Gaga has acted in the movie "A Star Is Born." The movie "A Star Is Born" is a musical romantic drama film that stars Lady Gaga and Bradley Cooper. Lady Gaga's performance in the movie earned her an Academy Award for Best Original Song for "Shallow."</td></tr>
</table>

Figure 20: Examples of refined responses generated by different methods on TopiOCQA. Fragments strongly aligned with the reference passage are highlighted in blue, while fragments with weaker connections (e.g., off-topic elements referring to previous conversation topics) are marked in red.