EVO-Reranker: Continuous Reranker Evolution under Multi-Round Feedback

Anonymous ACL submission

Abstract

Effective reranking is crucial for RAG performance, yet the continuous evolution of retrieval distributions necessitates multi-round optimization. We formulate this challenge under two scenarios: (1) **Full Feedback**, where iterative updates cause catastrophic forgetting of historical preferences; and (2) **No Feedback**, where the reranker suffers from a "dual-role" dilemma by acting as both learner and supervisor, leading to error propagation. We introduce **EVO-Reranker** to ensure safe continuous evolution. It employs **ReplayDPO** for feedback-rich environments to stabilize historical memory, and **CautiousDPO** for feedback-absent settings, leveraging confidence gating and multi-LLM consensus to ensure supervision quality. Extensive evaluations across authentic (e.g., BioASQ) and simulated shifts demonstrate that EVO-Reranker effectively averts performance collapse, translating to superior downstream generation on HotpotQA and MultiRC.

1 Introduction

Retrieval-Augmented Generation (RAG) grounds large language models (LLMs) in external evidence for knowledge-intensive tasks (Izacard and Grave, 2021; Guu et al., 2020). However, retrieval pipelines often return noisy or misleading candidate documents, which can degrade downstream reasoning and induce hallucinations (Shuster et al., 2021; Ji et al., 2023). Consequently, rerankers are crucial in RAG systems, directly controlling the context quality and heavily influencing the final generation.

Unlike static benchmarks, real-world RAG systems operate in dynamic environments where retrieval corpora and user queries continuously evolve. This non-stationary nature necessitates a multi-round continuous evolution paradigm, requiring the reranker to adapt to emerging domains across successive deployment rounds. In this study, we characterize such evolution based on the availability of expert annotations, formulating two realistic settings: **Full-Feedback**, where high-quality labels are continuously provided, and **No-Feedback**, where only initial annotations are available for subsequent rounds.

Adapting static preference optimization to multi-round environments exposes fundamental flaws. Under the full-feedback setting (continuous expert annotations), standard iterative DPO causes catastrophic forgetting, as the model overfits to shifting distributions and unlearns historical preference boundaries. Conversely, under the prevalent no-feedback setting (only initial annotations), naive self-training forces the reranker into a severe **"dual-role" dilemma**: it must act simultaneously as the optimization target and the sole supervisor. Relying entirely on self-generated pseudo-labels causes biased judgments to rapidly reinforce themselves, inevitably culminating in error accumulation and performance collapse. Alternatively, delegating the entire pseudo-label construction process to powerful external LLMs would incur prohibitive computational costs and massive resource waste.

To overcome these limitations, we propose **EVO-Reranker** (Figure 1), a unified framework providing structural guardrails for safe continuous evolution. For full feedback, **ReplayDPO** mitigates forgetting by replacing standard cumulative training with an error-aware historical replay mechanism that anchors hard-negative boundaries. For no feedback, **CautiousDPO** breaks the dual-role dilemma via cost-aware confidence gating. To balance supervision quality and computational expense, the model autonomously trusts its highly confident predictions, routing only uncertain samples to a multi-LLM consensus module to safely halt self-reinforcing noise. This mechanism significantly reduces resource consumption while en-

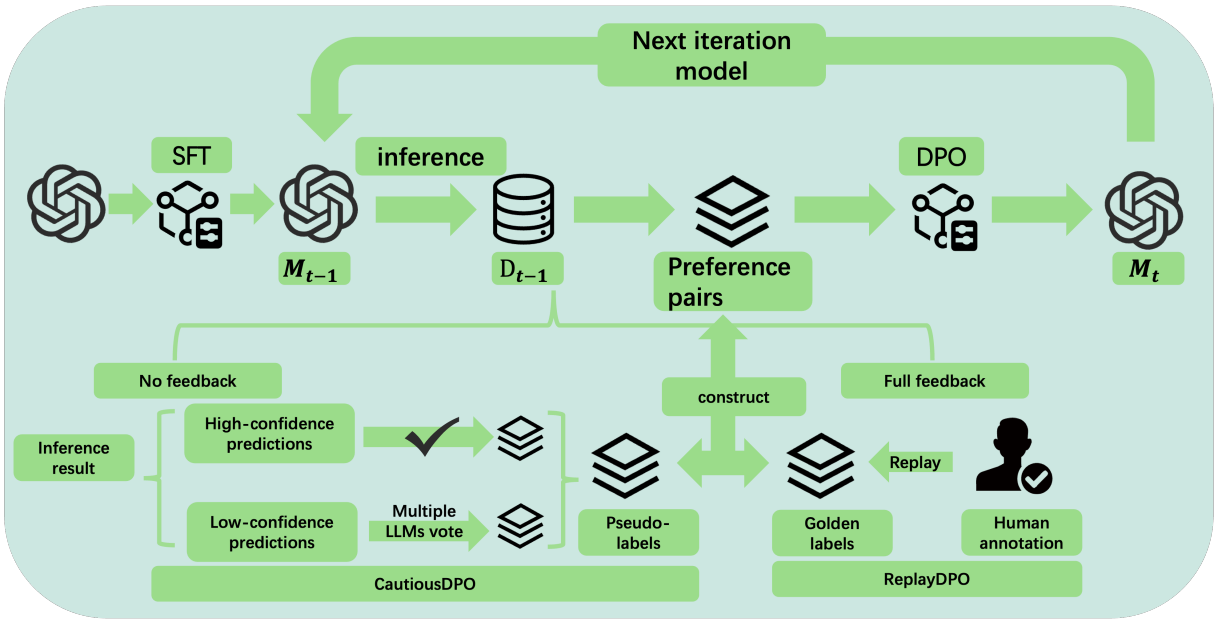

Figure 1: Overview of Multi-round Reranker Self-Evolution under Different Feedback Settings

abling highly effective and stable self-evolution.

Extensive experiments across six benchmarks demonstrate our framework's superiority. It achieves massive gains on authentic multi-round environments with genuine distribution shifts (e.g., BioASQ) and successfully halts the performance collapse inherent in naive self-training. Crucially, these sustained reranking improvements translate directly into significant gains in end-to-end downstream RAG answer generation.

Our main contributions are summarized as follows:

- We pioneer the **multi-round continuous reranker evolution** problem for dynamic RAG systems. We systematically dissect two scenarios: the **full-feedback setting**, which suffers from catastrophic forgetting, and the **no-feedback setting**, which faces the "dual-role" dilemma and self-training collapse.

- We propose the **EVO-Reranker** framework to provide safety guardrails. We introduce **ReplayDPO** with error-aware replay for full feedback, and **CautiousDPO** with cost-aware confidence gating and multi-LLM consensus for no feedback.

- Extensive evaluations demonstrate that EVO-Reranker averts the performance collapse of naive baselines, yields exceptional improvements in authentic shifting environments (up to +32.09% F1 on BioASQ), and consistently boosts **end-to-end generation quality**.

2 Related Work

2.1 Retrieval-Augmented Generation and Rerankers

Retrieval-Augmented Generation (RAG) is a standard paradigm for knowledge-intensive NLP tasks, where external documents are retrieved to support factual grounding and reasoning (Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022). In this pipeline, the quality of retrieved evidence is crucial for downstream generation.

Early RAG systems primarily rely on dense retrievers (Karpukhin et al., 2020), which often return noisy or misleading candidates. Rerankers therefore play a critical role by refining retrieval results and selecting highly relevant documents for generation (Miutra and Craswell, 2018; Lewis et al., 2020). Compared to retrievers, rerankers directly model fine-grained query–document relevance and have been shown to substantially improve generation quality.

Traditional rerankers are typically cross-encoder models based on BERT or T5 (Nogueira and Cho, 2019), which may struggle with complex semantic relevance and cross-domain generalization. Recent work shows that large language models exhibit strong ranking capabilities in both point-wise and comparative settings (Qin et al., 2024; Sun et al., 2023), making them

promising rerankers for RAG systems. Different from these approaches, we study reranker training in a realistic multi-round setting with evolving retrieval distributions and limited or diminishing human feedback.

2.2 Learning with Human and AI Feedback

Learning from feedback is a central paradigm for improving and aligning language models. Reinforcement Learning from Human Feedback (RLHF) leverages human preference annotations to guide model optimization, typically by training a reward model and optimizing the policy via reinforcement learning algorithms such as PPO (Schulman et al., 2017). Representative approaches include preference learning methods proposed by Ziegler et al. (2019), Stiennon et al. (2020), Ouyang et al. (2022), and Bai et al. (2022). More recent methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplify this pipeline by directly optimizing preferences without an explicit reward model, but still rely on high-quality human feedback.

However, in many realistic settings, especially under multi-round or continually evolving data distributions, obtaining accurate human preference annotations at every round is costly or infeasible. This has motivated growing interest in learning from model-generated feedback, often referred to as reinforcement learning from AI feedback (RLAIF) or self-training. Typical strategies include pseudo-labeling (Xie et al., 2020; Huang et al., 2023) and iterative self-refinement (Lyu et al., 2019). While these methods can reduce annotation cost, they are known to suffer from error accumulation and instability in multi-round learning, where early mistakes may be amplified over time. Similar to prior work on learning from AI feedback, we leverage model-generated signals for optimization, but focus on structuring and filtering them to enable stable preference learning across multiple rounds when continuous human feedback is unavailable.

3 Preliminaries

3.1 Point-wise Reranker and Preference Construction

Given a query q and a set of retrieved candidate documents $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, a point-wise reranker M_t evaluates each candidate independently to determine whether it contains information relevant to q (Nogueira and Cho, 2019; Guu et al., 2020). Formally, the reranker outputs a binary relevance decision:

$$M_t(q, s_i) \in \{\texttt{true}, \texttt{false}\}. \tag{1}$$

In preference-based optimization frameworks such as DPO, these point-wise judgments are subsequently used to construct pairwise preferences $(s_{\text{win}}, s_{\text{lose}})$, where a relevant document is preferred over an irrelevant one for the same query.

3.2 Multi-Round Learning Setting

We study a multi-round learning setting that reflects the continual operation of real-world RAG systems, where new query and document candidates are continuously introduced, causing distribution shifts over time.

At each round $t = 1, \ldots, T$, the system observes a newly retrieved batch:

$$\mathcal{D}_t = \{(q_i^{(t)}, \mathcal{S}_i^{(t)})\}_{i=1}^{N_t}, \tag{2}$$

where $q_i^{(t)}$ denotes a query issued at round t, and $\mathcal{S}_i^{(t)}$ is the set of retrieved candidate documents.

We denote the reranker available at the start of round t as M_{t-1}. The initial reranker M_0 is trained via supervised fine-tuning on a static labeled dataset. The multi-round continuous evolution proceeds sequentially through two phases at each round t:

(1) Inference and Evaluation: M_{t-1} is applied to the newly retrieved data $\mathcal{D}_t$ to produce relevance predictions. This phase evaluates how well the historical model M_{t-1} generalizes to the new distribution.

(2) Optimization and Update: The predictions or external feedback on $\mathcal{D}_t$ provide training signals to update the reranker from M_{t-1} to M_t. In standard baselines, M_t is trained cumulatively on the union of all available data up to round t (i.e., $\bigcup_{i=1}^{t} \mathcal{D}_i$)using SFT or DPO.

This sequential protocol formally captures the **dual-role dilemma**: at round t, M_{t-1} is both evaluated as the target model and utilized to construct the preference environment for training M_t. Consequently, explicitly modeling the stability of this update process is crucial for long-term RAG deployments.

4 Methodology

We formulate multi-round reranker optimization as a continual learning process. Rather than learn-

ing in a single static round, the model is repeatedly updated. Formally, learning at round t optimizes:

$$\max_{M_t} \mathbb{E}_{(q,s)\sim\mathcal{D}_t}[R_t(q,s)] \text{ s.t. } \mathrm{Dist}(M_t, M_{t-1}) \leq \epsilon, \tag{3}$$

where R_t denotes round-dependent relevance preferences induced by the current retrieval distribution, and M_{t-1} represents the reference reranker from the previous round to constrain updates for stability.

As established, the evolving reranker faces a **dual-role dilemma**, risking either catastrophic forgetting or self-training collapse. To mitigate these risks, we propose the **EVO-Reranker** framework, instantiating tailored safety guardrails for both full-feedback and no-feedback scenarios.

4.1 ReplayDPO: Error-Aware Historical Replay under Full Feedback

Under the full-feedback setting, expert relevance annotations are continuously available. Standard cumulative DPO often ignores historical boundaries, leading to oscillatory updates and catastrophic forgetting as distributions shift. To stabilize training, **ReplayDPO** maintains a historical replay buffer. At round t, the training set combines new and historical data:

$$\mathcal{D}_t^{\text{train}} = \alpha \cdot \mathcal{D}_t^{\text{new}} + (1-\alpha) \cdot \mathcal{D}_{1:t-1}^{\text{replay}}, \tag{4}$$

where $\mathcal{D}_t^{\text{new}}$ denotes preference pairs constructed from expert-labeled data at round t, and $\mathcal{D}_{1:t-1}^{\text{replay}}$ represents the preference set sampled from previous rounds.

Error-Aware Sampling Strategy: To elevate the replay mechanism beyond simple uniform data mixing, ReplayDPO employs an error-aware sampling strategy. Motivated by authentic multi-round refinement processes, we prioritize sampling hard-negative historical pairs. Specifically, we target preference pairs from past rounds where the current model M_{t-1} exhibits low prediction confidence or incorrect relevance ordering. This targeted replay acts as a strong anchor signal. It explicitly preserves previously aligned decision boundaries for confusing samples, enabling smooth reranker evolution while preventing the model from over-fitting to the current round's localized distribution.

4.2 CautiousDPO: Cost-Aware Confidence Gating and Consensus under No Feedback

In the no-feedback setting, the reranker must construct pseudo-preferences entirely from unlabeled retrieval data. Blindly trusting self-generated labels reinforces early errors, inevitably causing performance collapse. Conversely, full external multi-LLM calibration is computationally prohibitive.

To balance supervision quality and computational cost, we propose **CautiousDPO**, which adopts a **cost-aware confidence gating** mechanism coupled with heterogeneous consensus. The procedure operates systematically at each round:

(1) Confidence Estimation: For a newly retrieved batch, the current reranker M_{t-1} scores each candidate and outputs a generation probability. These probabilities serve as the confidence estimate for each prediction.

(2) Cost-Aware Gating: We establish a confidence threshold τ (e.g., selecting the top 70% most confident samples). Predictions exceeding this threshold are deemed reliable and are directly accepted as pseudo-labels. This step ensures that the majority of easy or familiar queries are processed with zero additional API cost.

(3) Multi-LLM Consensus Calibration: The remaining low-confidence cases, which are the primary catalysts for error accumulation, are routed to a consensus module. This module queries a panel of heterogeneous external LLM judges. The final pseudo-label for these uncertain samples is determined by a majority vote among the external judges and the current reranker.

By strategically restricting external calibration to this highly uncertain subset, CautiousDPO actively halts self-reinforcing noise, outputting high-quality preferences to optimize M_t with an optimal cost-performance trade-off.

5 Experiments

5.1 Datasets and Multi-Round Setup

We evaluate our framework on six benchmark datasets that span diverse real-world RAG scenarios, including general factual verification (**FEVER**, **ANLI**), multi-hop reading comprehension (**HotpotQA**, **MultiRC**), and specialized biomedical retrieval (**PubMedQA**, **BioASQ Synergy**). To rigorously evaluate continuous reranker evolution, all datasets are structured under a unified multi-round protocol. It is crucial to distin-

guish between two types of multi-round environments in our setup:

Authentic distribution shifts are evaluated using BioASQ Synergy as a genuine multi-round feedback environment. In this task, ranked documents are evaluated by expert assessors over four sequential rounds, where the specificity and long-tail nature of the queries naturally increase across rounds.

Simulated multi-round environments are constructed for the remaining five datasets to rigorously mimic real-world shifts. Rather than using trivial random sampling, we partition the data by clustering query text embeddings and assigning distinct semantic topics to sequential rounds (see Appendix B). This guarantees that the model continuously encounters novel query types, testing its ability to adapt without forgetting.

A consistent data scale is enforced across all environments to ensure fair comparison, with each round containing exactly 2,000 preference pairs. For standard cumulative baselines (Multi-round SFT and DPO), the model is trained on the expanding union of all accumulated data. In contrast, our proposed ReplayDPO maintains a fixed batch scale through targeted historical sampling.

5.2 Evaluation Metrics

We evaluate our framework on two dimensions: reranking quality and end-to-end downstream generation.

Reranking Quality: We align metrics with the inherent task formulations. For classification-centric tasks involving direct claim verification or binary true/false judgments (**FEVER**, **ANLI**, **PubMedQA**), we report Accuracy. For question-answering tasks requiring relevant evidence selection from large candidate pools (**MultiRC**, **HotpotQA**, **BioASQ**), we report the F1 score to account for class imbalance.

Downstream Generation: To verify whether improved reranking benefits the final RAG output, we feed the reranked documents into a fixed LLaMA-3 8B generator. We then evaluate the resulting answer quality using Exact Match (EM) and F1 on **MultiRC** and **HotpotQA**.

5.3 Compared Models

Implementation details are provided in Appendix A.1. We compare our framework against four categories of baselines:

Retrieval-based Rerankers: We evaluate **BGE-Reranker** (Li et al., 2023; Chen et al., 2024) (embedding similarity) and **MonoT5-3B** (Nogueira et al., 2020) (generative seq2seq scoring).

LLM Baselines and Judges: We evaluate the base **LLaMA-3 8B** (AI@Meta, 2024) and the instruction-tuned **ChatQA-1.5 8B** (Liu et al., 2024). For CautiousDPO's external judges, we utilize **DeepSeek-V3** (DeepSeek-AI, 2024) and **GPT-4o** (Hurst et al., 2024). Our preliminary tests confirm these external judges maintain high reliability (averaging 68% to 71% accuracy) even on the hardest low-confidence samples.

Static Baseline: SFT (LLaMA-3 8B) trained exclusively on the initial human-labeled dataset, representing standard deployment without continuous evolution.

Multi-round Baselines: Under full feedback, we evaluate **Multi-round SFT** and **Multi-round DPO** (Tu et al., 2025), which train cumulatively using gold labels at each round. Under the no-feedback setting, we introduce **Naive Multi-round Self-DPO** as a critical comparative baseline. It continuously trains on its own pseudo-labels without any confidence gating or external calibration, explicitly demonstrating the performance collapse caused by the dual-role dilemma.

5.4 Main Results

Table 1 reports the final reranking performance at Round 4. Overall, EVO-Reranker consistently dominates. CautiousDPO successfully prevents self-training collapse under no-feedback settings, while ReplayDPO achieves state-of-the-art results under full-feedback settings, yielding massive gains on the challenging BioASQ dataset. Specific findings are detailed below.

CautiousDPO ensures safe evolution under no-feedback settings. In the absence of expert feedback, blindly relying on self-generated pseudo-labels leads to severe error accumulation. As shown in Table 1, the Naive Multi-round Self-DPO baseline degrades significantly, falling behind the static SFT baseline on complex tasks like ANLI and BioASQ. In contrast, CautiousDPO effectively halts this performance collapse. By cautiously constructing preference signals, CautiousDPO attains 96.38% accuracy on FEVER and 81.39% F1 on HotpotQA, consistently outperforming the static SFT model and closely approaching

Table 1: Main results on reranking quality after multi-round continuous evolution (Round 4).

Method	FEVER (Acc %)	ANLI (Acc %)	PubMedQA (Acc %)	MultiRC (F1 %)	HotpotQA (F1 %)	BioASQ (F1 %)
Retrieval Baselines						
BGE-Reranker (2.72B)	76.38	36.63	79.40	60.95	53.96	29.79
MonoT5 (3B)	60.41	34.88	85.40	15.10	66.18	40.80
LLM Baselines						
LLaMA-3 (8B)	77.84	42.63	91.60	52.69	56.89	38.08
ChatQA-1.5 (8B)	69.83	41.62	49.80	61.39	48.44	41.43
DeepSeek-V3 (671B)	93.12	65.87	95.77	85.96	66.44	37.18
GPT-4o	90.92	63.45	97.05	86.95	68.98	41.85
Static Baseline (No Evolution)						
SFT (8B)	96.10	62.25	98.80	87.70	81.12	42.92
No-Feedback Setting (Self-Evolution)						
Naive Multi-round Self-DPO (8B)	94.65	57.80	96.50	84.20	78.40	38.15
CautiousDPO (Ours)	**96.38**	**66.85**	**99.45**	**88.10**	**81.39**	**44.62**
Full-Feedback Setting (Expert Supervision)						
Multi-round SFT (8B)	96.74	70.13	98.80	88.63	80.86	40.46
Multi-round DPO (8B)	96.64	72.62	99.45	88.46	81.93	44.76
ReplayDPO (Ours)	**96.78**	**72.88**	**99.55**	**88.65**	**82.55**	**76.85**

the upper bounds trained with expert annotations (i.e., full-feedback settings). These results demonstrate that CautiousDPO enables reliable and safe self-evolving reranker optimization.

ReplayDPO achieves strong gains under full-feedback settings. When full expert feedback is continuously available, ReplayDPO consistently outperforms standard Multi-round DPO and Multi-round SFT across the datasets. Standard multi-round training often suffers from oscillatory updates and forgets previously learned boundaries as the retrieval distribution shifts. By effectively replaying historical preferences, ReplayDPO anchors the model and mitigates catastrophic forgetting. For example, ReplayDPO achieves 72.88% accuracy on ANLI and a remarkable 76.85% F1 on BioASQ, illustrating a clear advantage over standard multi-round methods. BioASQ exhibits true sequence dependence, and the hard-negative nature of historical rounds is particularly important for defining complex medical boundaries. This indicates that explicitly modeling historical preference data is crucial for stable preference optimization in multi-round retrieval pipelines.

5.5 Ablation Study

To verify the contributions of our core components, we conduct an ablation study at evolution Round 4 (Table 2). Removing either mechanism significantly degrades performance.

Multi-LLM calibration is essential for robust no-feedback evolution. In the absence of continuous expert feedback, relying solely on internal confidence filtering proves insufficient. Without external consensus, the model suffers notable performance drops on complex tasks, such as ANLI (from 66.85% to 61.20%) and BioASQ (from 44.62% to 40.55%). Although this self-filtering variant still outperforms the Naive Self-DPO baseline (e.g., 61.20% vs. 57.80% on ANLI), the substantial gap to the full CautiousDPO proves that external calibration is absolutely essential to correct hard errors rather than merely ignoring them. A detailed empirical breakdown of the pseudo-label accuracy provided by individual external judges is presented in Appendix B.3.

Confidence gating dually preserves task alignment and resource efficiency. The necessity of selective gating becomes evident when querying external LLMs for all candidates. Such an "all-query" strategy not only incurs prohibitive computational costs but also actively degrades performance, with ANLI accuracy dropping from 66.85% to 63.50%. We attribute this to the fact that while external LLMs are powerful zero-shot reasoners, they lack the fine-grained task alignment embedded in the locally supervised reranker. Overriding high-confidence local predictions with zero-shot external judgments introduces domain-shift noise. Thus, confidence gating serves a dual

Table 2: Ablation study of CautiousDPO. Values represent the performance (Accuracy or F1 %) at the final evolution round (Round 4).

Method	FEVER	ANLI	PubMedQA	MultiRC	HotpotQA	BioASQ
CautiousDPO (Full)	**96.38**	**66.85**	**99.45**	**88.10**	**81.39**	**44.62**
w/o Confidence Gating	95.82	63.50	98.45	86.60	80.20	42.10
w/o Multi-LLM Calibration	95.21	61.20	97.60	85.45	79.50	40.55

Table 3: End-to-end RAG answer generation performance at the final evolution round (Round 4).

Reranker Model	MultiRC		HotpotQA	
	EM	F1	EM	F1
Static Baseline				
SFT (8B)	0.4210	0.7450	0.2225	0.4523
No-Feedback Setting				
Naive Self-DPO (8B)	0.3850	0.7020	0.1980	0.4150
CautiousDPO (Ours)	**0.4430**	**0.7680**	**0.3096**	**0.5465**
Full-Feedback Setting				
Multi-round SFT (8B)	0.4350	0.7590	0.2806	0.5444
Multi-round DPO (8B)	0.4520	0.7740	0.3161	0.5627
ReplayDPO (Ours)	**0.4680**	**0.7910**	**0.3290**	**0.5713**

purpose: it acts as a filter to retain task-specific expertise and as a cost-aware controller to minimize API overhead.

5.6 Impact of Reranking on Answer Generation

To assess end-to-end generation impact, we feed top documents from final-round rerankers into a fixed LLaMA-3 8B generator. Evaluated via EM and F1 on MultiRC and HotpotQA, Table 3 demonstrates that evolved reranker quality critically determines downstream RAG performance.

CautiousDPO averts downstream performance collapse. Under no-feedback settings, Naive Self-DPO accumulates pseudo-label noise that causes downstream QA performance to fall significantly below the static SFT baseline (e.g., F1 drops from 45.23% to 41.50% on HotpotQA). Conversely, CautiousDPO maintains high-quality evidence selection, yielding stable and substantial gains (54.65% F1 on HotpotQA) without expert supervision. This proves that robust reranking is a prerequisite for reliable long-term RAG.

ReplayDPO maximizes the generative upper bound. In the full-feedback setting, ReplayDPO consistently achieves the best overall generation performance. By mitigating catastrophic forgetting during reranking, it ensures the generator receives faithful and comprehensive context, reaching an impressive 32.90% EM and 57.13% F1 on HotpotQA. This indicates that preserving historical preference boundaries is essential for maximizing the reasoning potential of downstream generators.

5.7 Effect of Replay Ratio

Table 4 presents the sensitivity analysis of the replay ratio α on the ANLI dataset. The results reveal a clear inverted-U trend: extreme settings cause either catastrophic forgetting of historical boundaries ($\alpha = 1.0$) or structural over-anchoring to past data that hinders new adaptation ($\alpha \leq 0.4$). Empirically, moderate ratios strike the optimal balance, with $\alpha = 0.7$ yielding the most stable continuous evolution and the highest final accuracy (72.88%).

Table 4: Effect of Replay Ratio α on the ANLI dataset.

Alpha (α)	Round 2	Round 3	Round 4
0.1	67.85	68.20	68.50
0.2	68.40	69.55	69.80
0.3	69.15	70.80	70.50
0.4	70.50	71.65	71.10
0.5	**71.25**	72.88	72.75
0.6	69.87	72.12	72.25
0.7	70.25	**74.12**	**72.88**
0.8	69.37	72.88	72.88
0.9	69.50	71.88	72.75
1.0 (No Replay)	69.50	72.50	72.62

5.8 Preventing Performance Collapse

To visualize how EVO-Reranker prevents error accumulation, Figure 2 plots the performance trajectory on the ANLI dataset. Starting from the identical Round 1 checkpoint, the **Naive Self-DPO** model shows a deceptive initial gain but quickly succumbs to self-reinforcing noise. By Round 4, its performance completely collapses to 57.80%, falling significantly below the **Static SFT** baseline.

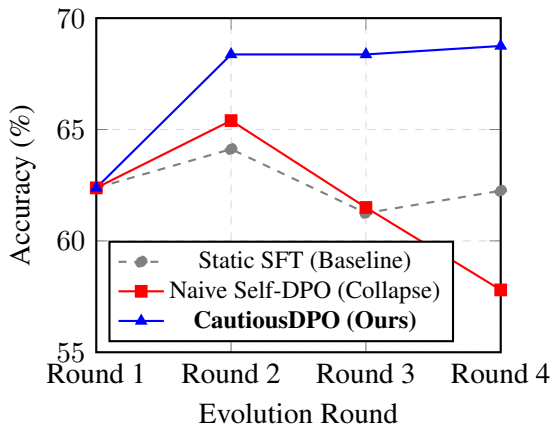

Figure 2: Evolution dynamics on the ANLI dataset across four rounds.

In contrast, **CautiousDPO** successfully halts this degradation. By filtering out destructive noise via confidence gating and multi-LLM consensus, it rapidly climbs from the initial baseline and maintains a robust upward trajectory, peaking at 68.75% in the final round. This dynamic explicitly confirms that safety guardrails are mandatory to overcome the dual-role dilemma and achieve stable continuous evolution.

5.9 Cost-Performance Trade-off Analysis

A potential concern with multi-LLM calibration is the computational and API cost. CautiousDPO effectively mitigates this through rank-based confidence gating. By trusting the top τ fraction of predictions, only the remaining $1-\tau$ fraction requires external judging.

To systematically determine the optimal threshold and quantify scalability, we define a **Normalized Trade-off Score** that equally weights performance retention and cost reduction (where cost is measured as a relative multiplier of the base local inference cost; detailed formulation is provided in Appendix B.1).

Our comprehensive sweep reveals that setting $\tau = 0.7$ achieves the optimal balance. Under this configuration, only 30% of the candidate instances are routed to external judges. Remarkably, this constrained intervention recovers over 40% of the performance gap between pure self-evolution (Naive Self-DPO) and full expert supervision. This confirms that CautiousDPO delivers substantial safety corrections at a highly manageable computational cost, making it exceptionally viable for real-world deployment.

5.10 Long-Horizon Dynamics

To examine long-horizon behavior, we extended CautiousDPO on ANLI through Round 6. Accuracy safely plateaus from 66.85% (Round 4) to 67.0% (Round 5) and 67.1% (Round 6), explicitly avoiding the degradation typical of naive self-training. This plateau is naturally driven by diminishing returns as pseudo-label quality stabilizes, alongside the intrinsic upper bound of self-evolution (ANLI's full-supervision ceiling is 72.88%). While authentic shifting environments like BioASQ sustain gains slightly longer, performance similarly tapers near the domain ceiling. These results confirm that EVO-Reranker guarantees a safe and stable plateau in extended deployments.

6 Conclusion

We study a practically important yet underexplored problem: continuous reranker evolution under multi-round distribution shifts. Unlike prior static methods, we recognize that dynamic RAG deployments expose fundamental flaws in standard optimization. Under **full-feedback** settings, models suffer from catastrophic forgetting as they overfit to shifting distributions. Conversely, under **no-feedback** settings, rerankers face a severe **"dual-role" dilemma** where they must act as both learner and supervisor, inevitably leading to self-reinforcing noise and performance collapse.

To ensure safe continuous evolution, we propose the **EVO-Reranker** framework to provide structural guardrails. For full feedback, we introduce **ReplayDPO**, which leverages an error-aware historical replay mechanism to anchor hard-negative boundaries and prevent forgetting. For the prevalent no-feedback scenarios, we propose **CautiousDPO**, which breaks the dual-role dilemma via cost-aware confidence gating and multi-LLM consensus, safely halting error accumulation while maintaining an optimal cost-performance trade-off.

Extensive evaluations demonstrate that EVO-Reranker averts the performance collapse of naive self-training and achieves massive retrieval gains on authentic shifting environments like BioASQ. Crucially, these sustained improvements translate directly into superior end-to-end RAG generation, establishing a highly reliable and cost-effective blueprint for real-world deployments.

Limitations

Despite its effectiveness, this work has several limitations. First, the proposed framework relies on multiple heterogeneous judges for preference calibration in the no-feedback setting, which introduces additional computational overhead and may limit applicability in strictly resource-constrained environments. Second, the stability of self-evolution depends on the quality of confidence estimation and inter-model agreement; inaccurate confidence signals or systematic bias among judges could reduce the reliability of constructed preferences, especially when transferring to unseen domains or model families. Finally, our study focuses on reranking within retrieval-augmented generation pipelines and evaluates primarily multi-hop reasoning and domain-specific benchmarks; extending multi-round preference optimization to more open-ended generation tasks or other components of RAG systems remains an open direction for future work.

Acknowledgments

We are grateful to our research collaborators and mentors for their valuable guidance and support throughout this project.

References

AI@Meta. 2024. Llama 3 model card.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1051–1068.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *Preprint*, arXiv:2312.15503.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*.

He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. 2019. Advances in neural information processing systems. *Advances in neural information processing systems*, 32.

Bhaskar Miutra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends$\hat{W}$ in Accounting*, 13(1):1–126.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 708–718.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Songjun Tu, Jiahao Lin, Xiangyu Tian, Qichao Zhang, Linjing Li, Yuqian Fu, Nan Xu, Wei He, Xiangyuan Lan, Dongmei Jiang, and 1 others. 2025. Enhancing llm reasoning with iterative dpo: A comprehensive empirical investigation. *arXiv preprint arXiv:2503.12854*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

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

A Appendix

This appendix provides additional details to complement the main paper, including ablation experiment results, dataset prompts, and example cases.

A.1 Implementation Details

All rerankers are based on LLaMA-3 8B with LoRA (rank 16). Models are initialized via SFT on 100K labeled instances and further optimized for 4 rounds using 2K instances per round with DPO. Experiments are conducted on NVIDIA A100 GPUs (80GB)

A.2 Detailed Analysis of Replay Ratio α

This section provides the comprehensive analysis of the replay ratio α, which governs the dynamic balance between newly collected preference data at the current round and historical preference data sampled from previous rounds. Table 4 reports the detailed performance trajectory of ReplayDPO on the ANLI dataset across different values of α.

The expanded results clearly demonstrate an inverted-U performance trend, confirming that moderate replay ratios consistently yield better and more stable performance than extreme settings at either end of the spectrum.

The danger of neglecting history ($\alpha \geq 0.8$): When α is too large (e.g., approaching 1.0), the training process relies almost exclusively on current-round data. While this allows fast adaptation, it makes the model highly susceptible to localized distribution shifts. The lack of historical anchoring leads to oscillatory updates and catastrophic forgetting, preventing the model from maintaining a stable performance gain across rounds.

The danger of over-anchoring ($\alpha \leq 0.4$): Conversely, when α is set too low, the model spends the majority of its optimization budget reviewing historical data. As shown in the top rows of Table 4, this over-anchoring creates structural rigidity. It severely restricts the model's capacity to adapt to newly emerging retrieval candidates and novel query types, resulting in a suboptimal and suppressed performance trajectory.

The optimal trade-off ($0.5 \leq \alpha \leq 0.7$): Empirically, values of α in the range of 0.5 to

0.7 achieve the best trade-off between continuous adaptation and historical stability. In particular, $\alpha = 0.7$ yields the highest peak accuracy at Round 3 (74.12%) and maintains superior stability through Round 4 (72.88%). These comprehensive results validate our core design principle: stable multi-round optimization requires a calibrated balance between exploring new preference environments and replaying historically corrected errors.

B Construction Details of Simulated Distribution Shifts

In Section 5.1, we introduced a semantic clustering strategy to simulate authentic multi-round distribution shifts for the five non-sequential datasets (FEVER, ANLI, PubMedQA, MultiRC, and HotpotQA). To ensure reproducibility and strict semantic boundaries between evolution rounds, we detail the construction pipeline below.

1. General Clustering Pipeline Instead of randomly partitioning the datasets—which fails to emulate real-world domain adaptation—we strictly segregated the data based on query semantics. For each dataset, we extracted the queries (or claims/hypotheses) and encoded them into dense vector representations using the pre-trained `BGE-base-en-v1.5` embedding model. We then applied the K-Means clustering algorithm on these embeddings to partition the data into $K = 5$ distinct semantic clusters (one for the initial SFT round, and four for the subsequent evolution rounds). To align with our controlled experimental setup, we randomly sampled exactly 2,000 preference pairs from each designated cluster to form the training batch for each respective round.

2. Dataset-Specific Semantic Shifts Because each dataset evaluates different capabilities, the embedding-based clustering naturally captured different types of structural and semantic shifts. We analyze the qualitative meaning of the shifting clusters for each dataset as follows:

- **FEVER (Fact Verification):** The queries in FEVER span diverse Wikipedia entities. Our clustering naturally grouped the claims by macroscopic topics. As the rounds progress, the reranker must evolve from verifying claims about *Entertainment and Pop Culture* (Round 1) to completely disjoint domains such as *Historical Events* (Round 2), *Geographic Entities* (Round 3), and *Science/Technology* (Round 4).

- **PubMedQA (Biomedical Retrieval):** Given its specialized medical nature, the clustering captured shifts across different medical subdisciplines. The multi-round sequence forces the model to continuously adapt to new medical vocabulary, shifting from *Pharmacology and Drug Interactions* to *Oncology*, and later to *Infectious Diseases*.

- **HotpotQA (Multi-hop Reasoning):** For HotpotQA, semantic clustering primarily captured the structural topology of the multi-hop questions. The distribution shift rigorously tests reasoning adaptation: earlier rounds might heavily feature "Bridge/Chaining" questions (e.g., finding the missing entity that connects two facts), while later rounds abruptly shift to "Comparison/Intersection" questions (e.g., comparing attributes like age or location between two entities).

- **MultiRC (Reading Comprehension across Genres):** MultiRC contains source texts from vastly different genres. The embedding clusters effectively segregated the queries based on these underlying source domains. Consequently, the evolving reranker must dynamically adjust its relevance boundaries from formal *News Articles* to *Fictional Stories*, and finally to *Historical Text* spans.

- **ANLI (Adversarial Inference):** ANLI is uniquely constructed via adversarial rounds. Our clustering further amplified the difficulty by grouping premise-hypothesis pairs based on distinct adversarial reasoning phenomena. During evolution, the reranker faces continuous shifts across different linguistic traps, moving from *Lexical Inference/Synonymy* to more complex *Numerical Reasoning* and *Coreference Resolution* traps.

By enforcing these strict semantic boundaries across rounds, our simulated environment rigorously prevents the model from relying on static lexical overlap, explicitly testing the core motivation of our EVO-Reranker framework.

B.1 Detailed Cost-Performance Trade-off Analysis

This section provides the complete data and formulation for the cost-performance trade-off analysis discussed in Section 5.9.

In the CautiousDPO framework, confidence is rank-based. The top τ fraction of samples are treated as high-confidence and accepted autonomously. Consequently, the external judging ratio is exactly $1 - \tau$.

We quantify the computational overhead using a **Relative Cost Multiplier**. Let the base cost of evaluating a candidate using the local 8B reranker be $1.0\times$. For the uncertain $1-\tau$ fraction routed to the external multi-LLM consensus module, an additional API overhead penalty is incurred. Given the pricing and latency of querying massive models (e.g., GPT-4o and DeepSeek-V3), we empirically estimate this external penalty to be $5.0\times$ the base cost. Thus, the expected relative cost per sample is modeled as a function of τ:

$$\text{Cost}(\tau) = 1.0 + 5.0 \times (1 - \tau) \tag{5}$$

As τ decreases, more samples are sent to the external multi-LLM panel, linearly increasing the overall relative cost but generally improving the quality of the pseudo-labels and the resulting accuracy.

To systematically quantify this balance, we define the **Normalized Trade-off Score** as follows:

$$\begin{aligned}\text{Score}(\tau) = \lambda &\frac{\text{Acc}(\tau) - \text{Acc}_{\min}}{\text{Acc}_{\max} - \text{Acc}_{\min}} \\ &+ (1-\lambda)\frac{\text{Cost}_{\max} - \text{Cost}(\tau)}{\text{Cost}_{\max} - \text{Cost}_{\min}},\end{aligned} \tag{6}$$

where $\text{Acc}(\tau)$ represents the reranking accuracy at threshold τ, and $\text{Cost}(\tau)$ is the corresponding relative inference overhead derived above. The terms with $\max$ and $\min$ subscripts denote the empirical upper and lower bounds observed during the threshold sweep.

Table 5 presents the empirical sweep of $\tau \in [0.0, 0.9]$ on the ANLI dataset at Round 4. The Normalized Score is calculated using $\lambda = 0.5$, which assigns equal weight to normalized accuracy retention and normalized cost reduction.

The extreme cases illustrate the boundary conditions. At $\tau = 0.0$, the system pays the maximum expected cost ($6.0\times$) to achieve the highest accuracy (72.80%), yielding a normalized score of 0.0 due to extreme cost inefficiency. Conversely, at $\tau = 0.9$, the system minimizes cost ($1.5\times$) but suffers severe performance degradation (63.80%), also yielding a score of 0.0. The configuration $\tau = 0.7$ emerges as the mathematically optimal point. It strategically allocates external LLM resources to the 30% most uncertain samples, successfully preventing self-training collapse while preserving 70% of the API budget.

Table 5: Cost-performance trade-off across different confidence thresholds (τ) on the ANLI dataset (Round 4). A higher Score indicates a better balance between accuracy and computational cost.

τ	Acc (%)	Relative Cost ($\times$)	Score
0.0	72.80	6.00	0.000
0.1	72.30	5.50	0.056
0.2	71.50	5.00	0.078
0.3	70.40	4.50	0.067
0.4	69.10	4.00	0.033
0.5	67.80	3.50	0.000
0.6	67.10	3.00	0.033
0.7	**66.85**	**2.50**	**0.106**
0.8	65.40	2.00	0.067
0.9	63.80	1.50	0.000

B.2 Detailed Analysis of Evolution Trajectories and Collapse

This section provides a deeper mechanistic analysis of the multi-round evolution dynamics depicted in Figure 2.

The Mechanics of Collapse: For the Naive Self-DPO baseline, the absence of an external calibration mechanism leaves the model entirely exposed to the dual-role dilemma. At Round 2, the model manages a slight improvement (from 62.38% to 65.40%) because the initial pseudo-labels generated by the strong SFT model still retain reasonable quality. However, as the model updates based on its own increasingly biased predictions, the signal-to-noise ratio in the training data drops rapidly. By Round 4, this vicious cycle drives the accuracy down to 57.80%, rendering the self-evolved model practically unusable compared to the static baseline.

The Mechanics of Stable Evolution and Plateau: CautiousDPO effectively breaks this destructive cycle. The confidence gating ensures that the model only reinforces itself on high-quality internal signals, while the multi-LLM consensus corrects the hard-negative samples that typically cause distribution drift. Consequently, CautiousDPO experiences a sharp performance surge at Round 2 (68.37%) and demonstrates a highly sta-

ble trajectory, culminating at 68.75% at Round 4.

This sustained stability reflects a natural plateau characteristic of safe self-evolution. Once the pseudo-label quality stabilizes and the evolving model approaches the intrinsic task ceiling achievable without continuous human supervision, the performance safely levels off. Unlike naive methods that crash under accumulated noise, CautiousDPO safely absorbs ongoing distribution shifts, proving its viability for long-horizon deployment.

B.3 Reliability of External LLM Judges on Hard Samples

In the CautiousDPO framework, external LLM judges are not queried for all candidate documents. Instead, due to the confidence gating mechanism, they are exclusively evaluated on the hardest, lowest-confidence subset of the data (i.e., the subset where the local reranker's confidence falls below the threshold τ). These ambiguous or adversarial samples inherently suppress the absolute accuracy scores of even the most capable frontier models.

Table 6 presents the pseudo-label accuracy of the heterogeneous LLM panel on this specific low-confidence subset of the ANLI dataset across the progressive evolution rounds.

Table 6: Pseudo-label accuracy (%) of individual external LLM judges and the aggregated consensus on the low-confidence subset of the ANLI dataset across evolution rounds. The heterogeneous consensus consistently outperforms individual models.

External LLM Judge	R1	R2	R3
Claude 3.5 Sonnet	71.20	68.50	67.10
DeepSeek-R1	70.50	67.80	66.20
Qwen2.5-72B	68.40	65.60	64.00
Consensus	**73.10**	**70.40**	**68.90**

Temporal Dynamics and Consensus Benefit: We observe a gradual decline in the absolute accuracy of all individual models as the rounds progress. This is an expected phenomenon in continuous evolution: as the local reranker improves at handling standard queries, the remaining samples that fall below the confidence threshold become increasingly difficult and domain-shifted.

Despite this increasing difficulty, the results validate our design choice of a heterogeneous panel. While individual models exhibit performance fluctuations, they tend to make different types of errors. By aggregating their predictions via majority voting (Heterogeneous Consensus), the ensemble successfully mitigates individual biases, consistently outperforming the best single model by approximately 2% across all rounds. This robust, unified calibration signal acts as a critical safety net, preventing error accumulation as the distribution shifts.

B.4 Case Study

We present qualitative case studies to illustrate how improved multi-round reranking affects both evidence selection and downstream answer generation. These examples highlight the different strengths of ReplayDPO and CautiousDPO under full-feedback and no-feedback settings, as well as remaining limitations.

Case Study 1 focuses on *reranking quality itself*. It shows that ReplayDPO is able to identify more relevant evidence and make correct relevance judgments, while multi-round SFT is misled by noisy retrieval results. This demonstrates that the performance gains of ReplayDPO stem not only from better generation, but also from improved document-level relevance modeling.

Case Studies 2 (B.4) and 3 (B.4) further illustrate how improved reranking translates into more accurate answer generation. In Case Study 2, ReplayDPO selects the correct supporting evidence, enabling the generator to identify the major tenant correctly, whereas Multi-round DPO is misled by irrelevant documents. In Case Study 3, CautiousDPO produces the correct profession by relying on more reliable pseudo-preference signals, while SFT yields an incorrect abstraction due to noisy self-supervision.

Case Study 4 (B.4) highlights a remaining limitation. Although ReplayDPO retrieves relevant evidence containing the correct information, the generator fails to fully integrate multiple retrieved sentences and instead outputs a closely related but incorrect concept. This suggests that errors can still arise from imperfect evidence integration at the generation stage, even when retrieval quality is substantially improved.

Case Study 1

Question: "Which are the most common

psychiatric events associated with the consumption of cannabis?"
Content: "AIMS: We tested the age-varying associations of cannabis use (CU) frequency and disorder (CUD) with psychotic, depressive and anxiety symptoms in adolescent and adult sa"

True label: "False"

Multi-round SFT: "True" (wrong)
ReplayDPO: "False" (correct)

Case Study 2

Question: "Which Canadian corporation, founded in March, 1994, is a major tenant of the Station Mall in Sault Ste. Marie?"

Ground Truth: "Wal-Mart Canada Corp."

Multi-round DPO: "Canadian Tire. (Source: Wikipedia)" (wrong)
ReplayDPO:" Walmart Canada Corp. is the major tenant." (correct)

Case Study 3

Question: "What profession does Alex Beard and Tony Hall, Baron Hall of Birkenhead have in common?"

Ground Truth:"director"

sft: "They are both CEOs of a major arts institution." (wrong)
CautiousDPO: "They are both directors." (correct)

Case Study 4

Question: "Migene González-Wippler is leading expert on a religion that developed among which ethnic group ?"

Ground Truth: "West African descendants"

ReplayDPO: "Afro-Caribbean"

Relevant sentences: "Migene González-Wippler is a Puerto Rican new-age author and a leading expert on the Afro-Caribbean religion of Santería.", "Santería, also known as Regla de Ochá, La Regla de Ifá, or Lucumi, is an Afro-American religion of Caribbean origin that developed in the Spanish Empire among West African descendants."

B.5 Dataset Prompts

Table 7 lists the prompts used for each dataset in our experiments.

Table 7: Prompt templates for each dataset. Curly braces like {query} or {claim} indicate placeholders for input values.

Dataset	**Prompt Template**
ANLI	Given a premise and a hypothesis, determine their relationship and explain your reasoning. First state the relationship (entailment, neutral, or contradiction). Premise: {premise} Hypothesis: {hypothesis} Relationship:
FEVER	You are a fact verification expert. Analyze the relationship between the claim and evidence. Claim: {claim} Evidence: {evidence} Based on the evidence, the claim is: - SUPPORTS if the evidence confirms the claim - REFUTES if the evidence contradicts the claim - NOT ENOUGH INFO if the evidence is insufficient Your response must be exactly one word: SUPPORTS, REFUTES, or NOT ENOUGH INFO.
HotpotQA	Determine whether the following sentence provides information relevant to answering the question. Question: {question} Sentence: {sentence} Answer (only True or False):
MultiRC	Based on the given passage and question, determine whether the provided answer is correct. If correct, answer 'true'; if incorrect, answer 'false'. Passage: {passage} Question: {question} Answer:
PubMedQA	Answer the medical research question based on the provided scientific context. Provide a 'yes' or 'no' answer. Question: {question} Scientific Context: {context} Answer:
BioASQ	Determine whether the given query and content are relevant. If so, answer 'true'; if not, answer 'false'. Query: {question} Content: {content} Relevance: