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

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) grounds large language models in external knowledge, making effective reranking essential, as irrelevant or noisy documents can impair downstream reasoning. In realistic deployments, evolving retrieval pipelines continuously generate new query–document candidates, naturally leading to a multi-round optimization problem for rerankers. However, effectively leveraging multi-round feedback is challenging, as initial labeled data is typically followed by unlabeled or weakly supervised data in later rounds, hindering sustained improvement. In this work, we study EVO-Reranker optimization under two realistic supervision scenarios. We consider two feedback settings. Under full feedback, where preference supervision is continuously available across rounds, we propose **ReplayDPO**. Under the more practical no-feedback setting, where only initial supervision is provided, we introduce **CautiousDPO**. Experiments on six benchmarks reveal that replaying historical preferences effectively mitigates catastrophic forgetting in multi-round optimization, leading to more stable and consistently superior reranking performance under full feedback. Moreover, by cautiously constructing preference signals from unlabeled data, CautiousDPO enables reliable self-evolution without expert supervision, substantially narrowing the performance gap to full-feedback settings. These results show that **EVO-Reranker** provides a unified framework for continuous reranker evolution, remaining effective across both full-feedback and feedback-limited settings.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm for enhancing large language models (LLMs) on knowledge-intensive tasks such as open-domain question answering, fact verification, and scientific reasoning (Izacard and Grave, 2021; Guu et al., 2020; Lewis et al., 2020). By conditioning generation on retrieved external documents, RAG systems ground model outputs in factual evidence. However, their effectiveness critically depends on the quality of documents retained for generation.

In practice, retrieval pipelines typically return a large set of candidate documents, many of which are irrelevant, misleading, or redundant. Feeding such noisy evidence into the generator can significantly degrade downstream reasoning and induce hallucinations (Shuster et al., 2021; Ji et al., 2023). Moreover, due to the lost-in-the-middle effect, truly relevant evidence may be overlooked when buried among long retrieval lists (Liu et al., 2024a). Consequently, rerankers are a key component of RAG systems, as they directly control the documents provided to the generator and often influence performance more than retrieval.

Unlike static benchmarks, real-world RAG systems operate in dynamic environments. Retrieval models, corpora, and user queries continuously evolve, producing new query–document candidates over time and naturally giving rise to multi-round feedback, where each deployment round yields new data that could be used to improve relevance modeling.

While existing pipelines can train rerankers with supervised relevance labels or preference data, they face fundamental limitations in multi-round settings. First, high-quality human relevance annotations are expensive, domain-specific, and difficult to obtain repeatedly as data distributions evolve (Nguyen et al., 2016; Thakur et al., 2021). In practice, supervision is often available only for an initial static dataset, while later rounds consist almost entirely of unlabeled retrieval outputs. Second, naive self-training or LLM-based pseudo-labeling strategies are prone to error reinforcement across rounds, as inconsistent or biased automatic

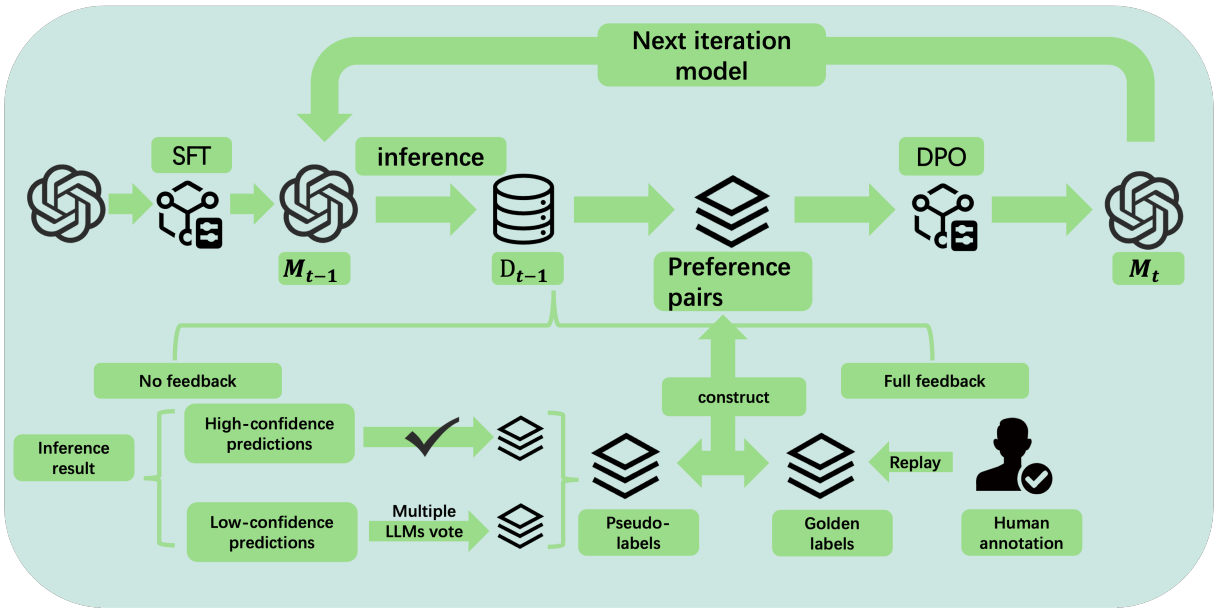

Figure 1: Overview of Multi-round Reranker Self-Evolution under Different Feedback Settings

judgments can accumulate over time and lead to performance collapse (Yarowsky, 1995; Zhou and Li, 2005; Wang et al., 2022; Liang et al., 2022). Together, these challenges make stable multi-round reranker optimization an open and underexplored problem.

In this work, we systematically study multi-round reranker optimization under two realistic supervision settings.

In the full-feedback setting, where accurate human preference feedback is available at each round, we propose **ReplayDPO**. ReplayDPO combines iterative preference optimization with preference replay to stabilize training under evolving retrieval distributions.

Beyond the full feedback setting, we consider the no-feedback setting, which is common in real deployments. In this setting, only an initial labeled dataset is available for supervised fine-tuning, while all subsequent rounds rely on unlabeled retrieval data. To address this challenge, we propose **CautiousDPO**, which carefully constructs reliable preference signals from unlabeled data by combining confidence-aware filtering with multi-LLM consensus. Extensive experiments across six benchmarks covering general factual verification, domain-specific factuality, and multi-hop reasoning show that ReplayDPO consistently outperforms standard multi-round training under full feedback, while CautiousDPO substantially outperforms strong retrieval baselines and single-round supervised rerankers under no feedback, significantly narrowing the gap to full-feedback upper bounds. These results show that explicitly modeling multi-round feedback dynamics is crucial for long-term reranker improvement in realistic RAG systems. Figure 1 provides an overview of our multi-round reranker evolution framework, which explicitly models evolving retrieval distributions and different feedback availability across rounds.

Our main contributions are summarized as follows:

- We introduce a realistic **multi-round reranker evolution setting** for RAG systems, where retrieval pipelines continuously generate new query–document candidates over time, while accurate human feedback is only available for an initial labeled dataset. This setting captures a key but underexplored challenge in real-world RAG deployment.

- Under the full-feedback setting, we propose **ReplayDPO**, which stabilizes iterative preference optimization through historical preference replay. ReplayDPO consistently outperforms multi-round SFT and DPO, achieving improvements on ANLI (up to 1.75% accuracy), HotpotQA (up to 2.6% F1), and especially BioASQ, a challenging domain-specific benchmark, with gains of up to 19.0% accuracy and 32.09% F1, while remaining stable across training rounds.

- Under the more practical no-feedback setting,

we propose **CautiousDPO**, which cautiously constructs reliable preference signals from unlabeled retrieval data via confidence-aware filtering and multi-LLM consensus. Without any additional human supervision, CautiousDPO achieves strong and consistent improvements across tasks, attaining 96.38% accuracy on FEVER and 81.39% F1 on HotpotQA, and substantially narrowing the gap to full-feedback training.

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 Using LLM

Given a query q and a set of retrieved candidate documents

$$S = \{s_1, s_2, \ldots, s_n\}, \tag{1}$$

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:

$$M_t(q, s_i) \in \{\texttt{true}, \texttt{false}\}. \tag{2}$$

3.2 Multi-Round Learning Setting

We study a **multi-round** learning setting that reflects the continual operation of real-world RAG systems, where retrieval results evolve over time and new query–document candidates are continuously introduced.

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

$$\mathcal{D}_t = \{(q_i^{(t)}, \mathcal{S}_i^{(t)})\}_{i=1}^{N_t}, \tag{3}$$

where $q_i^{(t)}$ denotes a query issued at round t, and $\mathcal{S}_i^{(t)} = \{s_{i1}^{(t)}, \ldots, s_{in}^{(t)}\}$ is the set of retrieved candidate documents.

We denote the reranker used in round t as M_{t-1}. Multi-round learning proceeds sequentially: at round t, M_{t-1} is applied to the newly retrieved data $\mathcal{D}_t$ to produce relevance predictions and pairwise preferences. These predictions both evaluate the generalization of M_{t-1} on unseen data and provide training signals to update the reranker to M_t, which is trained on data up to round t and implicitly evaluated on the next-round distribution $\mathcal{D}_{t+1}$. The initial reranker M_0 trained via supervised fine-tuning on a static labeled dataset.

This unified inference-driven protocol captures the non-stationary nature of realistic RAG pipelines, where each reranker must both adapt to and be evaluated on continuously shifting retrieval distributions.

4 Methodology

We study multi-round preference-based reranker optimization for Retrieval-Augmented Generation (RAG) systems, where retrieval results and relevance distributions evolve over time. Rather than learning a reranker in a single static round, we consider a continual setting in which the model is repeatedly updated using preference signals constructed at each round.

Formally, learning at round t can be viewed as optimizing:

$$\max_{M_t} \mathbb{E}_{(q,s)\sim\mathcal{D}_t}[R_t(q,s)] \text{ s.t. } \mathrm{Dist}(M_t, M_{t-1}) \leq \epsilon. \tag{4}$$

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, which constrains updates for training stability.

Under this formulation, preference signals may originate from different supervision sources. We consider two practical settings commonly encountered in real-world systems: **(i) full feedback**, where expert relevance annotations are available at each round, and **(ii) no feedback**, where no human supervision is provided. For each setting, we design a corresponding optimization strategy that enables stable multi-round reranker evolution.

4.1 ReplayDPO: Preference Optimization with Historical Replay under Full Feedback

Under the full-feedback setting, expert relevance annotations are available for retrieval results at every round. This enables a supervised form of multi-round reranker evolution, where the model can be continuously refined as new feedback becomes available. A straightforward strategy is to apply DPO independently at each round. However, this ignores preference signals accumulated in earlier rounds and can lead to unstable or oscillatory updates as retrieval distributions evolve.

To address this issue, we propose **ReplayDPO**, a multi-round preference optimization method that stabilizes training by replaying historical preference data. At round t, the training set is constructed as:

$$\mathcal{D}_t^{\text{train}} = \alpha \cdot \mathcal{D}_t^{\text{new}} + (1-\alpha) \cdot \mathcal{D}_{1:t-1}^{\text{train}}, \tag{5}$$

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

The first round is initialized using preferences derived from the model's initial inference results. To further stabilize early-stage optimization, we optionally replay a small subset of preference data converted from the initial SFT training corpus, which serves as an anchor signal for relevance alignment. From the second round onward, ReplayDPO balances newly observed preferences with historical signals, enabling smooth reranker evolution while preserving previously aligned behavior.

4.2 CautiousDPO: Confidence-Aware and Consensus-Based Preference Optimization under No Feedback

In the no-feedback setting, no expert relevance annotations are available beyond the initial round. The reranker must therefore evolve across rounds by constructing preference signals from unlabeled retrieval data.

CautiousDPO enables stable multi-round evolution by adopting a conservative self-supervision strategy. At each round, high-confidence predictions from the current reranker are directly used as

Algorithm 1 Multi-Round Preference Optimization with **CautiousDPO** (No Feedback)

Input: Initial reranker M_0; confidence threshold τ; external LLM judges $\{\mathrm{LLM}_k\}_{k=1}^{K}$; number of rounds T
Output: Final reranker M_T
for $t = 1$ **to** T **do**
  Receive newly retrieved candidates $\mathcal{D}_t = \{(q_i^{(t)}, \mathcal{S}_i^{(t)})\}_{i=1}^{N_t}$
  Initialize preference set $\mathcal{P}_t \leftarrow \emptyset$
  for each query $q_i^{(t)}$ with candidates $\mathcal{S}_i^{(t)}$ **do**
    Compute relevance scores and confidence estimates: $\{(s, \hat{y}_s, c_s)\} \leftarrow M_{t-1}(q_i^{(t)}, \mathcal{S}_i^{(t)})$
    for each candidate $s \in \mathcal{S}_i^{(t)}$ **do**
      if $c_s \geq \tau$ **then**
        Accept prediction $\hat{y}_s$ as reliable
      else
        Query external judges and set $\hat{y}_s \leftarrow \mathrm{Consensus}(\mathrm{LLM}_1, \ldots, \mathrm{LLM}_K, M_{t-1})$
      end if
    end for
    Construct preference pairs from predicted labels $\mathcal{P}_t \leftarrow \mathcal{P}_t \cup \{(q_i^{(t)}, s^+) \succ (q_i^{(t)}, s^-)\}$
  end for
  Update reranker M_t by optimizing DPO on $\mathcal{P}_t$
end for

preference signals, while low-confidence cases are resolved through consensus among heterogeneous judges, including strong LLMs and the previous-round reranker. This design mitigates error accumulation by avoiding over-reliance on uncertain pseudo labels during self-evolution.

The resulting pseudo-preferences are optimized using DPO to update the reranker for the next round. The complete procedure is summarized in Algorithm 1.

We term this approach **CautiousDPO** to emphasize its cautious evolution mechanism, where preference signals are adopted only when supported by high confidence or cross-model agreement.

5 Experiments

5.1 Datasets

We evaluate our methods on six benchmark datasets widely used for document-level relevance modeling in RAG, covering factual verification, multi-hop reasoning, and domain-specific retrieval. All datasets are organized under a unified multi-round evaluation protocol to reflect continual retrieval and feedback in realistic deployments.

BioASQ Synergy (Nentidis et al., 2024) naturally follows a multi-round feedback setting. In the BioASQ Synergy task, ranked documents are submitted over four rounds, with expert assessors providing relevance annotations at each round, forming an authentic multi-round supervision scenario. For the remaining datasets, including **FEVER** (Thorne et al., 2018), **ANLI** (Nie et al., 2020), **HotpotQA** (Yang et al., 2018), **MultiRC** (Khashabi et al., 2018), and **PubMedQA** (Jin et al., 2019), we simulate multi-round learning by partitioning the development sets into four sequential rounds. In each round, new query–document candidates are introduced, and relevance labels are revealed only for the current round, thereby modeling evolving retrieval distributions and delayed supervision. Together, these datasets enable systematic evaluation of multi-round reranker evolution across diverse tasks and feedback conditions.

5.2 Evaluation Metrics

We follow the standard evaluation protocols of each dataset and primarily evaluate document relevance filtering performance.

ANLI, FEVER, and PubMedQA are treated as classification tasks, and we report classification accuracy. For MultiRC, BioASQ, and HotpotQA, which involve evidence selection from multiple candidates, we report both accuracy and F1 to account for class imbalance.

In addition, to assess the impact of reranking on downstream generation, we use a fixed LLaMA-3 8B generator across all methods and evaluate answer generation quality using Exact Match (EM) and F1.

5.3 Compared Models

We compare our methods with representative baselines that cover retrieval-based rerankers, LLM-based relevance modeling, supervised training, and multi-round full-feedback upper bounds. Implementation details are provided in Section A.1.

Retrieval-based rerankers. We evaluate two strong rerankers: **BGE-Reranker** (Li et al., 2023; Chen et al., 2024), which ranks candidates via non-generative embedding similarity, and **MonoT5-3B** (Nogueira et al., 2020), a generative seq2seq

Table 1: Main results on general factual verification. Accuracy (%) on FEVER and ANLI is reported across multiple training rounds; numbers in parentheses indicate gains over baselines (CautiousDPO vs. SFT, ReplayDPO vs. multi-round DPO).

Method	FEVER (%)				ANLI (%)			
	R1	R2	R3	R4	R1	R2	R3	R4
Retrieval Baselines								
BGE-Reranker (2.72B)	76.94	75.88	76.60	76.38	32.62	35.75	36.00	36.63
MonoT5 (3B)	60.56	60.88	59.81	60.41	35.50	33.00	33.50	34.88
LLM Baselines								
LLaMA-3 (8B)	77.70	77.74	77.24	77.84	41.38	45.62	44.75	42.63
ChatQA-1.5 (8B)	69.40	69.86	68.75	69.83	40.62	44.62	44.12	41.62
DeepSeek-V3 (671B)	93.40	93.34	93.34	93.12	64.00	66.25	64.50	65.87
GPT-4o	91.68	91.40	91.34	90.92	66.75	64.88	63.00	63.45
Fine-tuning Baselines								
SFT (8B)	96.64	96.66	95.98	96.10	62.38	64.12	61.25	62.25
CautiousDPO(8B)	–	96.82	96.22	96.38	–	68.37	68.37	66.75
		(+0.16)	(+0.24)	(+0.28)		(+4.25)	(+7.12)	(+4.50)
Multi-round SFT(8B)	–	97.06	96.52	96.74	–	69.87	71.63	70.13
Multi-round DPO(8B)	–	97.08	96.42	96.64	–	69.50	72.50	72.62
ReplayDPO(8B)	–	**97.20**	**96.56**	**96.78**	–	**71.25**	**72.88**	**72.88**
		(+0.12)	(+0.14)	(+0.14)		(+1.75)	(+0.38)	(+0.26)

Table 2: Main Results on Multi-hop Reasoning (Rounds 2-4). We report accuracy and F1 (%) on MultiRC and HotpotQA; numbers in parentheses indicate gains over baselines (CautiousDPO vs. SFT, ReplayDPO vs. multi-round DPO).

Method	**MultiRC (%)**						**HotpotQA (%)**					
	R2		R3		R4		R2		R3		R4	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Retrieval Baselines												
BGE-Reranker (2.72B)	47.07	57.60	48.75	59.27	50.57	60.95	40.85	52.72	40.52	53.21	40.93	53.96
MonoT5 (3B)	54.33	16.57	51.50	14.37	51.39	15.10	72.09	66.63	70.99	66.43	71.75	66.18
LLM Baselines												
LLaMA-3 (8B)	70.40	59.44	68.69	53.63	66.31	52.69	69.33	56.78	69.74	56.89	69.12	56.89
ChatQA-1.5 (8B)	53.01	57.96	54.40	60.36	57.83	61.39	72.14	50.35	71.30	49.46	70.74	48.44
DeepSeek-V3 (671B)	89.61	86.93	86.79	83.63	88.09	85.96	77.43	66.08	77.62	66.08	77.37	66.44
GPT-4o	88.79	86.43	87.54	84.94	88.66	86.95	76.69	68.75	77.94	69.39	77.99	68.98
Fine-tuning Baselines												
SFT (8B)	89.53	88.01	88.29	86.92	88.58	87.70	83.38	79.81	82.96	80.05	84.28	81.12
CautiousDPO (8B)	89.61	88.27	88.37	87.03	88.83	88.10	84.93	81.14	84.69	81.85	84.79	81.39
	(+0.08)	(+0.26)	(+0.08)	(+0.11)	(+0.25)	(+0.40)	(+1.55)	(+1.33)	(+1.73)	(+1.80)	(+0.51)	(+0.27)
Multi-round SFT (8B)	89.94	88.36	88.62	86.99	89.48	88.63	82.04	79.45	81.34	79.42	83.32	80.86
Multi-round DPO (8B)	89.78	88.21	89.62	87.88	89.23	88.46	85.07	80.79	**86.61**	**82.79**	86.00	**81.93**
ReplayDPO (8B)	**90.35**	**88.67**	**89.62**	**87.76**	**89.56**	**88.65**	**87.46**	**83.36**	85.26	81.11	**86.21**	81.62
	(+0.57)	(+0.46)	(+0.00)	(-0.12)	(+0.33)	(+0.19)	(+2.39)	(+2.57)	(-1.35)	(-1.68)	(+0.21)	(-0.31)

model for relevance scoring.

LLM-based baselines. We evaluate the relevance modeling capability of large language models, including the base **LLaMA-3 8B** (AI@Meta, 2024), the instruction-tuned **ChatQA-1.5 (8B)** (Liu et al., 2024b), and strong external LLM judges, **DeepSeek-V3** (DeepSeek-AI, 2024) and **GPT-4o** (Hurst et al., 2024).

Single-round supervised baseline. We include **SFT (LLaMA-3 8B)**, trained on a static human-labeled dataset, representing standard supervised relevance learning without multi-round evolution.

Multi-round methods. Our proposed methods operate in two supervision settings: **CautiousDPO**, which performs multi-round preference optimization without human feedback, and **ReplayDPO**, which leverages full human feedback with preference replay. To contextualize their performance, we additionally report two full-feedback upper bounds: **Multi-round SFT**, which applies standard SFT at each round, and **Multi-round DPO**, which performs DPO independently at each round

Table 3: Main Results on Domain-specific Factuality. We report accuracy and F1 (%) on PubMedQA and BioASQ across multiple training rounds; numbers in parentheses indicate gains over baselines (CautiousDPO vs. SFT, ReplayDPO vs. multi-round DPO).

Method	PubMedQA (%)				BioASQ (%)							
	R1	R2	R3	R4	R1		R2		R3		R4	
	Acc	Acc	Acc	Acc	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Retrieval Baselines												
BGE-Reranker (2.72B)	79.15	78.75	77.20	79.40	23.30	29.92	25.94	34.31	25.69	32.93	24.31	29.79
MonoT5 (3B)	85.40	86.35	85.10	85.40	59.01	39.08	60.18	43.06	61.16	42.40	63.62	40.80
LLM Baselines												
LLaMA-3 (8B)	90.75	89.45	91.20	91.60	47.92	39.86	49.06	41.49	49.78	40.46	51.28	38.08
ChatQA-1.5 (8B)	50.40	50.35	47.70	49.80	57.53	38.78	55.34	44.31	56.79	47.85	57.76	41.43
DeepSeek-V3 (671B)	95.25	95.00	95.40	95.77	46.74	38.47	47.99	39.88	48.79	39.10	50.60	37.18
GPT-4o	96.85	96.30	97.35	97.05	58.29	41.04	63.65	44.02	59.15	42.08	64.69	41.85
Fine-tuning Baselines												
SFT (8B)	98.25	98.05	98.65	98.80	62.91	46.06	63.20	47.25	63.17	45.51	64.53	42.92
CautiousDPO (8B)	–	99.05	99.25	99.45	–	–	69.60	48.94	69.49	47.07	70.85	44.62
		(+1.00)	(+0.60)	(+0.65)			(+6.40)	(+1.69)	(+6.32)	(+1.56)	(+6.32)	(+1.70)
Multi-round SFT (8B)	–	98.40	99.25	98.80	–	–	68.21	45.94	65.16	47.89	65.86	40.46
Multi-round DPO (8B)	–	**99.25**	99.40	99.45	–	–	77.44	50.11	70.12	47.13	70.95	44.76
ReplayDPO (8B)	**–**	98.95	**99.45**	**99.55**	**–**	**–**	**80.98**	**61.22**	**82.84**	**71.05**	**90.00**	**76.85**
		(-0.30)	(+0.05)	(+0.10)			(+3.54)	(+11.11)	(+12.72)	(+23.92)	(+19.05)	(+32.09)

Table 4: Ablation study across all datasets. Values are average accuracy (%) over rounds 2-4 for each configuration.

Method	**General Factual**		**Multi-hop Reasoning**		**Domain-specific**	
	FEVER	ANLI	HotpotQA	MultiRC	PubMedQA	BioASQ
CautiousDPO	96.47	67.83	84.80	88.94	99.25	69.98
w/o Confidence Gating	96.37	64.87	82.83	88.28	99.20	68.85
w/o Multi-LLM Calibration	96.33	62.25	82.90	88.25	99.12	65.62

using gold preferences (Tu et al., 2025).

5.4 Main Results

Tables 1, 2, and 3 report the main results under different feedback settings. Overall, CautiousDPO excels on reasoning-intensive and domain-specific tasks under no-feedback settings, while ReplayDPO consistently improves multi-round optimization, with especially large gains on BioASQ due to its high reliance on specialized biomedical knowledge. The specific findings are summarized as follows.

CautiousDPO shows Effectiveness under no-feedback settings In the absence of expert feedback, CautiousDPO consistently achieves strong performance across all datasets and remains highly competitive with methods trained under full-feedback supervision. For instance, CautiousDPO attains 96.38% accuracy on FEVER and 81.39% F1 on HotpotQA, closely matching or approaching the best results obtained with expert annotations. These results show that CautiousDPO's careful pseudo-labeling enables effective self-evolving reranker optimization.

ReplayDPO achieves strong gains under full-feedback settings. When full expert feedback is available, ReplayDPO consistently outperforms standard Multi-round DPO and Multi-round SFT on the majority of datasets. Overall, ReplayDPO demonstrates consistently strong performance under the full-feedback setting and, in most cases, surpasses other compared multi-round optimization methods. For example, ReplayDPO achieves 72.88 accuracy on ANLI and 76.85 F1 on BioASQ, illustrating its advantage over standard Multi-round DPO and SFT. This indicates that replaying historical preference data is crucial for stable and effective preference optimization in multi-round retrieval settings.

5.5 Replay Ratio Analysis in Multi-round Training

We analyze the effect of the replay ratio α in multi-round training and find that moderate replay yields more stable and effective performance; detailed results are provided in Section A.2.

Table 5: Experimental results on HotpotQA dataset across different rounds.

Model	Round 2		Round 3		Round 4	
	EM	F1	EM	F1	EM	F1
LLaMA-3	0.2264	0.4479	0.2134	0.4372	0.2225	0.4523
CautiousDPO	0.2954	0.5461	0.2889	0.5445	0.3096	0.5465
Multi-round SFT	0.3149	0.5548	0.2824	0.5531	0.2806	0.5444
Multi-round DPO	0.3182	0.5577	0.3214	0.5696	0.3161	0.5627
ReplayDPO	**0.3344**	**0.5632**	**0.3344**	**0.5896**	**0.3290**	**0.5713**

Table 6: Retrieval-augmented Answer Generation Results on HotpotQA across Evolution Rounds.

Method	**R1**	**R2**	**R3**
ANLI			
SFT-only	62.38	64.13	61.25
CautiousDPO	**69.50**	**70.63**	**69.87**
BioASQ			
SFT-only	71.74	69.24	68.68
CautiousDPO	**79.32**	**75.01**	**73.87**

5.6 Ablation Study

To verify the contributions of its components, we conduct an ablation study of CautiousDPO (Table 4). Removing either confidence gating or multi-LLM calibration consistently degrades performance across datasets. The effect is more pronounced on reasoning-intensive and domain-specific tasks. Removing multi-LLM calibration leads to notable drops on ANLI (67.83% $\rightarrow$ 62.25%) and BioASQ (69.98% $\rightarrow$ 65.62%), indicating the importance of consensus-based supervision under ambiguity. Disabling confidence gating also causes performance declines, though smaller. In contrast, gains on FEVER and PubMedQA are limited, likely because these benchmarks are relatively simple and already close to task saturation.

5.7 Impact of Reranking on Answer Generation

We evaluate end-to-end answer generation on HotpotQA using a fixed LLaMA-3 8B generator with documents selected by different rerankers (Table 5). Improved reranking consistently boosts EM and F1, demonstrating that stronger evidence selection directly benefits multi-hop generation. CautiousDPO provides stable gains under no-feedback settings (F1 $\approx$ 0.54–0.55), while ReplayDPO achieves the best overall performance with full feedback (F1 0.5896, EM 0.3344). These results show that progressive reranker evolution translates into clear end-to-end gains. Appendix A.3 presents case studies illustrating how higher-quality evidence leads to more faithful and complete answers.

5.8 Pseudo-label Quality Analysis

Table 7 shows that CautiousDPO consistently yields higher-quality pseudo labels than the SFT-only baseline across evolution rounds and datasets. Despite a gradual accuracy drop in later rounds due to increased retrieval difficulty, CautiousDPO maintains a clear margin, indicating improved calibration rather than error amplification. These results suggest that its gains under no-feedback settings are primarily driven by more reliable pseudo-preference signals.

6 Conclusion

We study a practically important yet underexplored problem: multi-round reranker evolution under limited or missing feedback. Unlike prior reranking and preference optimization methods that rely on single-round supervision or continuous expert annotations, this setting requires rerankers to improve progressively under distribution shifts and incomplete feedback. To address this challenge, we propose **Evo-Reranker**. In the no-feedback setting, we introduce **CautiousDPO**, which enables stable self-evolution by combining confidence-based filtering with consensus from heterogeneous judges. In the full-feedback setting, we further propose **ReplayDPO** to reuse historical preferences and alleviate catastrophic forgetting across rounds. Extensive experiments on multi-hop and domain-specific benchmarks demonstrate that our methods consistently improve reranking quality over rounds and that these gains translate into stronger downstream generation in end-to-end RAG pipelines, especially in long-horizon, feedback-constrained settings.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

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.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. 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 TrendsŴ in Accounting*, 13(1):1–126.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, and Georgios Paliouras. 2024. Overview of bioasq tasks 12b and synergy12 in clef2024. *Working Notes of CLEF*, 2024.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541.

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–5K instances per round with DPO ($\beta = 0.2$). Experiments are conducted on NVIDIA A100 GPUs (80GB)

A.2 Effect of Replay Ratio α

We analyze the effect of the replay ratio α, which controls the balance between newly collected preference data at the current round and historical preference data accumulated from previous rounds. Table 7 reports the performance of ReplayDPO on ANLI across evolution rounds under different values of α.

Overall, moderate replay ratios consistently yield better performance than extreme settings. When α is too large (e.g., $\alpha = 1.0$), training relies solely on current-round data, making the model more sensitive to distribution shifts and leading to less stable improvements across rounds. Conversely, overly small values of α limit the model's ability to adapt to newly emerging retrieval candidates.

Empirically, values of α in the range of 0.5–0.7 achieve the best trade-off between adaptation and stability. In particular, $\alpha = 0.7$ yields the highest accuracy at Round 3, while $\alpha = 0.5$ and $\alpha = 0.7$ consistently perform well across all rounds. These results validate the importance of balancing historical preference replay with newly collected feedback for stable multi-round optimization.

A.3 Case Study

We present qualitative case studies to illustrate how improved multi-round reranking affects both evidence selection and downstream answer generation.

Table 7: Retrieval-augmented Answer Generation Results on HotpotQA across Evolution Rounds.

Alpha	R2	R3	R4
	ANLI		
0.5	**71.25**	72.88	**72.88**
0.6	69.87	72.12	72.25
0.7	70.25	**74.12**	72.75
0.8	69.37	72.88	72.88
0.9	69.50	71.88	72.75
1.0	69.50	72.50	72.88

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 (A.3) and 3 (A.3) 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 (A.3) 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."

A.4 Ablation Experiment Results

Tables 8, 9, and 10 show the detailed results of each round of our ablation experiments. These data supplement the summarized results reported in the main paper.

A.5 Dataset Prompts

Table 11 lists the prompts used for each dataset in our experiments.

Table 8: Ablation study on General Factual Verification datasets (FEVER and ANLI). Results are reported as accuracy (%).

Method	**FEVER (%)**			**ANLI (%)**		
	R2	R3	R4	R2	R3	R4
CautiousDPO	96.82	96.22	96.38	68.37	68.37	66.75
w/o Multi-LLM Calibration	96.66	96.06	96.26	63.62	61.62	61.50
w/o Confidence Gating	96.72	96.14	96.32	65.50	65.00	64.12

Table 9: Ablation study on Multi-hop Reasoning datasets. Values shown are (Accuracy / F1) % averages for each round.

Method	**HotpotQA (Acc/F1)**			**MultiRC (Acc/F1)**		
	R2	R3	R4	R2	R3	R4
CautiousDPO	84.93/81.14	84.69/81.85	84.79/81.39	89.61/88.27	88.37/87.03	88.83/88.10
w/o Confidence Gating	83.33/80.19	82.91/80.60	82.26/79.68	89.12/87.78	88.21/87.02	87.52/87.04
w/o Multi-LLM Calibration	83.33/80.19	82.76/80.37	82.72/80.09	89.12/87.78	88.29/86.88	88.34/87.64

Table 10: Ablation study on Domain-specific Factuality datasets (PubMedQA and BioASQ)

Method	**PubMedQA (%)**			**BioASQ (%)**		
	R2	R3	R4	R2	R3	R4
CautiousDPO	99.05	99.25	99.45	69.60	69.49	70.85
w/o Confidence Gating	99.05	99.40	99.30	68.65	68.10	69.80
w/o Multi-LLM Calibration	99.05	99.05	99.25	69.78	63.05	64.04

Table 11: 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: