Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Lightweight Continuous-Regret Policy Distillation for Adaptive Scientific Evidence Retrieval

Xiao Liu
State Grid Sichuan Electric Power Information & Telecommunication Company
Yuhan Dai
State Grid Sichuan Electric Power Information & Telecommunication Company
Wenjie Chen
State Grid Sichuan Electric Power Information & Telecommunication Company
Yaomu Tan
State Grid Sichuan Electric Power Information & Telecommunication Company
Zhenjie Han
Sichuan Energy Internet Research Institute, Tsinghua University
Ziang Yao
yaoziang@tsinghua-eiri.org

Sichuan Energy Internet Research Institute, Tsinghua University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11430197/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Lightweight Continuous-Regret Policy Distillation for Adaptive Scientific Evidence Retrieval

Xiao Liu[1], Yuhan Dai[1], Wenjie Chen[1], Yaomu Tan[1], Zhenjie Han[2,3], Ziang Yao[2*]

[1]State Grid Sichuan Electric Power Information & Telecommunication Company, No. 16 Jinhui West 2nd Street, High-Tech Zone, Chengdu, 610041, Sichuan, China.
[2]Sichuan Energy Internet Research Institute, Tsinghua University, Building 5, Zone A, Tianfu New Economy Industrial Park, Chengdu Science City, Chengdu, 610213, Sichuan, China.
[3]Chongqing University, Chongqing, China.

*Corresponding author(s). E-mail(s): yaoziang@tsinghua-eiri.org;

Abstract

Scientific claims can be classified into two retrieval methods: one is where the names and thresholds in the evidence sentences are almost exactly repeated, and the other is a free paraphrasing without retaining keywords. There must be a fixed fusion rule to simultaneously meet these two requirements. We propose a Continuous Regret Policy Distillation (CRPD) strategy, which is a small control layer that combines the scores from the six fixed strategies in the retrieval combination learned for each query. The training process is conducted offline: each evaluated query receives scores under each strategy, the utility is transformed into the regret value for each action, and these results are predicted from the query and retrieval state signals through a shallow regressor ensemble model; no re-training of any retriever is required, nor is a language model run during query evaluation. On the SciFact dataset, CRPD achieves 0.7174 nDCG@10, improving RRF by 0.0276 (95% paired bootstrap confidence interval [0.0121, 0.0440], Holm correction $\boldsymbol{p = 0.0016}$), outperforming the independently trained strategy classifier and the adapted QDAP-S on both datasets, where the difference in ANTIQUE is significant after Holm correction ($\boldsymbol{p \leq 0.0008}$). Since the direct utility of matched capacity cannot be statistically distinguished from CRPD, these two continuous objectives share the same effective region. All five SciFact training partitions remain above the preset strongest fixed fusion threshold, with the RRF gain of the lowest sparse-dense overlap quartile reaching +0.0585. The cost of feature generation and median strategy prediction is 0.961 milliseconds per query. When a certain action dominates, such as in the ANTIQUE task, a strong fixed hybrid model is still the better choice.

Keywords: scientific evidence retrieval, adaptive retrieval, hybrid retrieval, policy distillation, continuous regret

1 Introduction

Some scientific claims resemble database queries: they list a drug, an enzyme, or a measurement condition, and the supporting sentences almost exactly repeat these keywords; for example, "mercaptopurine" and "TPMT" appear letter by letter on both sides of the query. Other claims fall into the problem of synonymous paraphrasing: statements about the behavior of biological materials may not be consistent with the surface form that

supports the abstract concept. Vocabulary retrieval is applicable to the first type of situation and has a lower cost; while dense encoders are suitable for the second type of situation. Neither of them can simultaneously meet these requirements, so the hybrid sparse-dense stack has become the default solution for scientific evidence search. However, the usual hybrid model adopts a single global fusion setting, which is not applicable to any pattern to which the input query belongs. What is lacking in this field is not another fusion rule; it is a method of selecting from the existing rules at each query time.

The selection of strategies for queries has been studied as a trade-off between efficiency and effectiveness [1], and adaptive fusion methods now adjust the balance of sparsity and density in each query based on retrieval confidence and further relevance signals [2]. Both types of methods have confirmed the value of query adaptation. However, they both guide the adapter by compressing decisions. For each training query, a "best method" label records which strategy won and discards the margin: confusing two almost equal strategies, their scores are the same as choosing a strategy with extremely poor performance. Additionally, a controller that is cheap enough to deploy should be able to read the complete action-utility table and make corresponding decisions.

We propose a Continuous Regret Policy Distillation (CRPD) method for scientific evidence retrieval. The training process is divided into three stages. Each training query first executes against a frozen combination of six fixed strategies: dictionary run, dense run, ranking-based fusion, and three fixed score mixture models, with the dense weights being 0.25, 0.50, and 0.75 respectively. The generated utility table, with each row corresponding to a query and each column corresponding to an operation, is then rewritten as the "regret value" for each query, that is, which operations are effective for a certain query in the supervision record and how much the wrong choice incurs a cost. Finally, through a set of shallow regressors, the regret value of each operation is predicted using the signals available during retrieval (such as query statistics, the shape and consistency of the two result lists). The soft strategy converts these six predicted regret values into the mixed weights of the frozen score list.

Since the retriever itself does not move, the entire learning system acts as a control layer, directly built on the running mixed stack. Its sole responsibility is to map the observable retrieval state to the operation preferences. No retrievers are retrained, nor is the index replaced; in the service stage, the relevance judgment and language model do not participate. Each component retriever retains all the validations it has passed.

Our contributions include: a supervision mechanism for maintaining the complete query-level utility table of the fixed six strategy combinations during the training process, instead of a single argmax label for each query. A lightweight controller (based on query statistics, score shape, cross-run overlap, and ranking-inconsistency shallow regressors), whose output directly reweights the frozen score list; a controlled evaluation that covers fixed fusion rules and non-training fusion rules, as well as independently trained strategy classifiers, and combines the adaptive QDAP-S model to achieve the goal of capacity matching and decoding control under paired multi-comparison statistics; in addition, a deployment-oriented reading: five-zone stability, adaptation-opportunity analysis, measured controller overhead, worked real cases, and the low-heterogeneity boundary at which a fixed mixture is still the better choice.

The evaluation revolves around three questions: whether the distillation of continuous regret can raise evidence retrieval above the fixed rules and the learned adaptive alternative methods; which component supports the observed behavior, namely the continuous multi-action supervision, the retrieval-state features, or the soft decoding; and under what retrieval conditions the adaptive strategy can obtain benefits, and, equally important in operation, when the fixed fusion rule should continue to be used.

2 Related Work

2.1 Sparse, Dense and Mixed Retrieval

Although BM25 [3] is a sparse method, it is still used as a standard term ranking method in many evidence retrieval processes. Dense retrieval takes a different approach: it does not rely on exact term overlap, but ranks candidates based on the similarity learned in the vector space, such as in Dual

Encoder Retrieval [4]. Learning-based sparse methods like SPLADE [5] are in between; they retain the inverted index while extending the document's lexical representation.

Unlike the aforementioned methods, the subsequent interactive model ColBERT [6] improves token-level scoring by expanding the index size; the encoder BGE-M3 [7] can fuse dense representations, learned sparse representations, and multi-vector patterns; the performance of the encoder was evaluated only in the dense mode.

Integrating the two ranking methods presents challenges: RRF [8] mixes the ranking lists but does not calibrate its scoring scale, while amplitude information can support score-based fusion under the assumption of a single global mixed weight [9]. Further controlled analysis shows that compared to RRF, its performance is more sensitive to normalization methods, the position of the mixed weights, and the degree of deviation between the test domain and the training domain.

2.2 Complementarity and Adaptive Selection

Re-training the retriever is one way to achieve complementarity. RoC [10] separates sparse retrieval errors from dense retrieval errors by adjusting the training objectives of the two branches; in contrast, this method takes the opposite approach: keeping the retriever unchanged and confining all learning processes to a lightweight strategy layer that only reads the ranking results and does not access other information. For indexed services that have been built, this design difference is of practical significance, because re-training the retriever with an encoder requires re-encoding the entire corpus; freezing the pairing does not require this operation.

The selector lies between the two schemes: in a budget-constrained situation, the classifier trained by Arabzadeh et al. [1] is responsible for determining whether to call the sparse branch, the dense branch, or both. QuDAR [2] combines the score difference with the relevance signal generated by a large language model to construct query-level fusion weights. In contrast, QDAP [11] only takes the query vector as input and obtains a single sparse-dense mixture coefficient through regression; based on this approach, we further adapted its lightweight design as QDAP-S, a direct learning baseline model.

Compared to the above systems, CRPD has three main differences: first, this method does not involve a language model; the action space consists of six predefined strategies instead of a single scalar coefficient or three-way calls; in addition, the teacher part is represented by an effectiveness table constructed through manual evaluation of queries, enabling students to observe the cost associated with each error operation, rather than relying solely on occasional heuristic methods.

From a system perspective, the cascade controller reaches similar conclusions: the workload allocation should be determined by the query; AgentIR [12] adjusts its cascade process based on the workload, while adaptive reordering [13] routes each request to the retriever or re-arranger based on expected utility. The difference of CRPD lies in its supervision stage, that is, for each judged training query, a score is given to each strategy in the combination; subsequently, the student is trained to fit a continuous regret vector.

2.3 Scientific Evidence Retrieval

In the verification of scientific literature facts, hybrid, multi-stage processing flows are increasingly being adopted; the shared task submission method [14] combines lexical and embedding-based evidence retrieval, while the climate science validation system [15] integrates multiple retrievers and reordering mechanisms. However, these methods do not aim to derive "regret teachers" from a fixed strategy set. This study addresses this deficiency. Adaptive RAG [16] proposes a related concept, where the depth of retrieval is regarded as a bandit decision learned through interaction. In contrast, CRPD, in the offline stage, considers the effectiveness matrix as deterministic and fully observable, and uses this as the basis for training a compact supervised student model.

2.4 Policy Distillation

The "policy distillation" discussed here refers to how the teacher conveys information to the students. Each time a query is processed, the hard label transmits only one bit of information, namely "this query belongs to action k". When two actions are almost equivalent, this label will flip randomly during the training rounds, thereby hiding the cost of each error. Our teacher retains the utility of every action, converts the columns into regret

values, and then softens the results into a target distribution suitable for regression by the gradient boosting ensemble model. This process does not require the retrieval stack to be differentiable end to end.

3 Continuous-Regret Policy Distillation

3.1 Strategy portfolio and offline utility

Let $\mathcal{S}$ represent the fixed set of retrieval strategies. Our combination includes six strategies: BM25, BGE-M3-DENSE, RRF, and the minimum-maximum score fusion strategy with weights of 0.25, 0.50, and 0.75 respectively. For each query q and strategy s, the offline teacher calculates:

$$\begin{aligned} U(q,s) = w_1\,&\mathrm{nDCG@10}(q,s) \\ &+ w_2\,\mathrm{Recall@10}(q,s) \\ &+ w_3\,\mathrm{Recall@100}(q,s). \end{aligned} \tag{1}$$

We set $(w_1, w_2, w_3) = (0.5, 0.2, 0.3)$, emphasizing the quality of early sorting, while maintaining shallow and broad evidence coverage through two recall terms. The combination and utility weights are fixed before testing.

Query oracles and action regret are defined as follows:

$$U^\star(q) = \max_{s\in\mathcal{S}} U(q,s), \tag{2}$$

$$\Delta(q,s) = U^\star(q) - U(q,s). \tag{3}$$

Regret $\Delta(q,s)$ measures the utility loss caused by choosing strategy s instead of the oracle operation. It distinguishes between near-optimal alternatives and costly decisions, providing richer supervision information compared to hard action labels.

3.2 Soft continuous-regret teacher

We convert these regrets into a soft target distribution:

$$p_T(s \mid q) = \frac{\exp[-\Delta(q,s)/\tau]}{\sum_{s'\in\mathcal{S}} \exp[-\Delta(q,s')/\tau]}, \tag{4}$$

The temperature parameter τ determines the teacher's preference for the optimal strategy. Since $\Delta(q,s) = U^\star(q) - U(q,s)$, this distribution is equivalent to $p_T(s \mid q) = \mathrm{softmax}(U(q,s)/\tau)$. The regret formula also provides a non-negative, query-related target for the student regressor. We compare its effect with a capacity-matching model trained directly based on utility.

3.3 Retrieval-time features

To predict the strategy performance for unseen queries, CRPD uses features extracted from queries and initial sparse and dense rankings. These features cover three aspects of retrieval behavior: query characteristics, score distribution, and consistency between retrievers.

Query characteristics include length, digit ratio, mean and maximum inverse document frequency (IDF), and rare word item ratio. For each retriever, score statistics include the highest score, the gaps between top-1 and top-2 and between top-1 and top-10, mean, standard deviation, and normalized entropy. Cross-retriever features include Top-100 Jaccard similarity, overlap coefficient, ratios of only sparse and only dense, Spearman rank correlation, and average normalized ranking difference, measuring the overlap and sorting consistency of candidate options.

We divide these features into four categories,

$$\mathbf{x}(q) = [\mathbf{x}_Q(q); \mathbf{x}_S(q); \mathbf{x}_D(q); \mathbf{x}_{SD}(q)], \tag{5}$$

Here, $\mathbf{x}_Q$ contains the vocabulary features at the query level, $\mathbf{x}_S$ and $\mathbf{x}_D$ summarize the sparse and dense rating distributions respectively, and $\mathbf{x}_{SD}$ reflects the consistency between their rankings. This structure also supports the ablation of feature groups.

The rating entropy is used to measure the degree to which the retriever concentrates its ratings on a few candidates. Given a rating vector $\mathbf{a} = (a_1, \ldots, a_k)$, we define $v_i = a_i - \min_j a_j$, and calculate $\pi_i = v_i / \sum_j v_j$, using a uniform distribution when all v_i are zero. The calculation of normalized entropy is as follows:

$$H_k(\mathbf{a}) = -\frac{1}{\log k} \sum_{i=1}^{k} \pi_i \log \pi_i. \tag{6}$$

A lower entropy value indicates a more concentrated rating, while a higher entropy value indicates

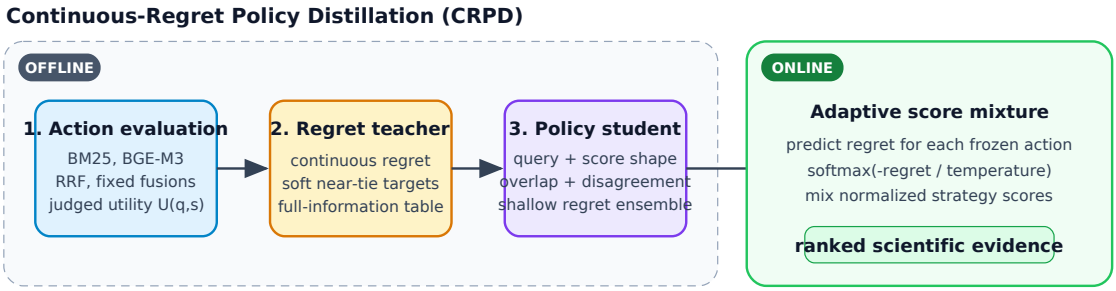

Fig. 1 CRPD follows three stages: full-information offline action evaluation, continuous regret prediction, and an adaptive score mixture. The online student uses retrieval-time features only.

a flatter distribution. We calculate these statistics for each retriever separately to retain its own rating scale.

The rating statistics describe the individual characteristics of each retriever. Cross-retriever features are used to capture whether sparse retrieval and dense retrieval tend to use the same evidence. We measure this consistency through candidate overlap and relative ranking order. At depth k:

$$J_k(q) = \frac{|L_S^k(q) \cap L_D^k(q)|}{|L_S^k(q) \cup L_D^k(q)|}, \tag{7}$$

$$G_k(q) = \frac{1}{|I_k(q)|} \sum_{d \in I_k(q)} \frac{|r_S(d) - r_D(d)|}{k}, \tag{8}$$

Here, $I_k(q)$ represents the shared candidate set, r_S and r_D represent the sparse and dense rankings respectively. The Jaccard overlap J_k is used to measure the shared candidates, while G_k reflects the inconsistency in their rankings. These features help distinguish the consistency of the candidate set from the differences in ranking order.

3.4 Regret student

For each strategy s, a group of shallow gradient boosting regressors is used to predict its action regret value. These ensemble members share the same feature structure and model family, but use different random seeds. Their average predicted value is:

$$\widehat{\Delta}(q,s) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\Delta}_m(q,s), \tag{9}$$

The predicted regret value is converted into strategy weights:

$$p_S(s \mid q) = \frac{\exp[-\widehat{\Delta}(q,s)/\tau]}{\sum_{s'} \exp[-\widehat{\Delta}(q,s')/\tau]}. \tag{10}$$

3.5 Adaptive score mixture

The resulting weights define the query adaptive hybrid model of six retrieval strategies. Let $\widetilde{f}_s(q,d)$ represent the query-level normalized score of document d under strategy s. The final ranking score is

$$F(q,d) = F_{\text{mix}}(q,d) = \sum_{s \in \mathcal{S}} p_S(s \mid q) \widetilde{f}_s(q,d). \tag{11}$$

Throughout the entire main evaluation process, this direct mixed configuration was adopted. Furthermore, an optional RRF rollback mechanism was implemented for the deployment diagnosis. This mechanism would be activated when the predicted regret margin, normalized by ensemble disagreement, fell below a certain threshold. During the development stage, a zero threshold was selected for both datasets, so this rollback mechanism was not enabled in all reported experiments.

3.6 Offline compilation and online complexity

The offline teacher evaluated each judged training query under each strategy, generating a complete utility matrix containing $|Q_{\text{tr}}||\mathcal{S}|$ query-strategy entries. The resulting regret values were used to guide the student regressor. In the deployment stage, the controller only acted on query features and retrieval outputs without performing relevance judgments.

Algorithm 1 CRPD inference

Require: query q, strategy runs $\{L_s(q)\}$, regret ensemble

1: extract query, score-shape, and cross-run features
2: predict $\widehat{\Delta}(q, s)$ for all s
3: compute soft weights $p_S(s \mid q)$
4: **return** weighted normalized mixture of $\{L_s(q)\}$

With feature dimension F, ensemble size M, and shallow tree depth T, after the component rankings became available, the online strategy evaluation required $\mathcal{O}(|\mathcal{S}|MTF)$ operations. In our implementation, this cost was relatively small compared to query encoding and vector search. Additional online computations included feature aggregation, regret value prediction, and weighted score fusion.

3.7 Scientific evidence application workflow

The complete workflow consists of five stages: scientific proposition, mixed retrieval, offline strategy evaluation, student learning based on regret, and adaptive evidence sorting. The sparse and dense retrievers are still in a frozen state.

4 Experimental Setup

4.1 Primary application and external stress test

We evaluated CRPD using SciFact and ANTIQUE. SciFact serves as our primary benchmark for verifying scientific statements [17], while ANTIQUE is used to test its generalization ability in non-factual question answering [18]. We adopted the standard training and test set division method for both datasets. We reserved 20% of the training queries for development, and then re-fitted the final model on the complete training set.

4.2 Retrievers and baselines

Our retrieval strategy combination consists of six methods: BM25, BGE-M3 dense retrieval, RRF ($c = 60$), and the minimum-maximum score fusion based on 0.25, 0.50, and 0.75 dense weights. We compared CRPD with these six strategies and three adaptive methods. The methods without training calculate the dense weights for each query based on the standardized sparse and dense Top-1–Top-10 score differences, and adjust their bias and scaling parameters on the development set.

In addition, we trained two alternative models for query adaptation. The gradient boosting strategy classifier uses the same features as CRPD to predict the utility-maximizing strategy. Its class probabilities determine the fusion weights. QDAP-S (adapted) uses the frozen BGE-M3 query embeddings to predict a probability distribution containing 101 discrete sparse-dense mixed weights, combining linear projection and one-dimensional smooth convolution [11]. Training uses cross-entropy loss and Wasserstein loss. Both models are learned based on the same composite utility calculated from the top-100 retrieval results.

We conducted four control experiments to test the role of each component. Direct utility regression uses the same architecture and training settings as regret value regression to predict $U(q, s)$. The global mixed average method averages the teacher weights in the training queries to test the advantage of adapting the weights to a single query. Hard decoding selects a single strategy from the predicted regret values to compare with soft fusion methods. Variants based solely on query features exclude retrieval-derived features and are used to measure their contribution.

4.3 Evaluation and statistics

We evaluated retrieval performance using nDCG@10, Recall@10, and Recall@100. The average strategy utility reflects the performance under the composite objective. The Oracle-gap closure measures how much available improvement space CRPD can recover through query adaptation. Tables 1 to 3 show the dataset statistics and computational costs.

For each dataset, we used 10,000 pairs of bootstrap resampled samples to estimate the two-sided confidence intervals. Holm correction was used to handle multiple comparisons. For methods a and b, we defined the difference for each query as:

$$\delta_i(a, b) = m_i(a) - m_i(b), \qquad \bar{\delta}(a, b) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \delta_i(a, b), \tag{12}$$

Where m_i is the evaluation metric of query i. We obtained the confidence interval by re-sampling the paired differences $\{\delta_i\}$. Given the ordered p-values $p_{(1)} \leq \cdots \leq p_{(M)}$, the value after Holm correction is:

$$p_{(j)}^{\text{Holm}} = \max_{1 \leq \ell \leq j} \min\bigl(1, (M - \ell + 1) p_{(\ell)}\bigr). \quad (13)$$

We define the oracle-gap closure as the ratio between the utility gain of CRPD over a fixed baseline and the corresponding oracle gap:

$$\text{GapClosure} = \frac{\overline{U}_{CRPD} - \overline{U}_b}{\overline{U}_{\text{oracle}} - \overline{U}_b}, \quad (14)$$

For each query, the oracle selects the strategy with the highest utility from the predefined portfolio. The oracle-gap closure thus represents the improvement degree of CRPD relative to the specific reference for that query.

4.4 Implementation, model selection, and reproducibility

All experiments used a maximum search depth of 100 and a strategy seed of 20260728. The search engine remained unchanged, and we selected the model settings on the development set. The retrieval pipeline follows reproducible sparse/dense experiment conventions supported by toolkits such as Pyserini [19]. To ensure reproducibility, we saved the TREC-style rankings, the metrics for each query, the teacher objectives, and the model prediction results.

5 Results

5.1 Overall effectiveness

The locked test results appear in Table 4. Among the single-retriever, fixed-fusion, and training-free adaptive methods listed there, CRPD attains the highest SciFact value at 0.7174 nDCG@10. RRF stands at 0.6897, the strongest prespecified fixed fusion at 0.7101, and training-free dynamic fusion at 0.7099. Against RRF the paired difference reaches +0.0276, with 95% CI [+0.0121, +0.0440] (raw $p = 0.0002$; Holm-adjusted $p = 0.0016$).

5.2 Policy utility and oracle gap

Utility 0.8512 marks the SciFact query oracle, 0.8052 the best static strategy, and 0.8123 CRPD. Working only from retrieval-time features at deployment, the distilled policy thereby closes 15.4% of the static-to-oracle utility gap. The complete six-action strategy portfolio, not nDCG@10 alone, enters this calculation.

5.3 Comparison with learned adaptive alternatives

Fixed fusion is a weaker test than the independently learned adaptive baselines. Across the two reported datasets, nDCG@10 for CRPD stays above both the policy classifier and adapted QDAP-S. SciFact gives CRPD 0.7174, against 0.7022 for the soft policy classifier and 0.7083 for QDAP-S, a paired gain of +0.0151 and +0.0090. Because their Holm-adjusted intervals include zero, the SciFact gains remain directionally consistent yet statistically unresolved. The same ordering does reach statistical resolution on ANTIQUE, where CRPD exceeds the soft policy classifier by +0.0157 (95% CI [0.0072, 0.0248], Holm-adjusted $p = 0.0008$) and QDAP-S by +0.0187 (95% CI [0.0103, 0.0274], Holm-adjusted $p \leq 0.0008$). Direction is therefore consistent across datasets, with a significant margin on the external stress test.

5.4 Continuous-target and policy controls

Target parameterization, policy decoding, and query-independent mixing are separated from the independently learned alternatives in Table 6. On SciFact the capacity-matched direct-utility student posts 0.7210 nDCG@10 and on ANTIQUE 0.3575, beside 0.7174 and 0.3592 for CRPD. Paired CRPD–utility intervals cross zero on both datasets, and the point-estimate ordering reverses between them. Both continuous targets consequently occupy one effectiveness region, which indicates comparable effectiveness across regret- and utility-centered parameterizations.

Adaptive continuous mixing also comes into sharper focus through the controls. The query-independent portfolio mixture trails CRPD on SciFact by +0.0113 (95% CI [0.0029, 0.0207], Holm-adjusted $p = 0.027$). Decoding the CRPD regret vector by hard selection reaches only 0.7055 on

Table 1 Dataset and evaluation scale. All reported queries have relevance judgments.

Dataset	Documents	Train queries	Test queries	Run depth
SCIFACT	5,183	809	300	100
ANTIQUE	403,666	2,426	200	100

Table 2 Measured first-stage query latency in milliseconds on the fixed experimental machine. Index construction is excluded.

	BM25		BGE-M3-DENSE	
Dataset	p50	p95	p50	p95
SCIFACT	0.12	0.24	13.25	15.50
ANTIQUE	0.88	10.83	92.05	101.68

Table 3 Online operation profile. CRPD adds no neural encoder beyond the component hybrid system.

Method	Sparse run	Dense run	Learned online control
BM25	yes	no	none
BGE-M3-DENSE	no	yes	none
RRF / fixed fusion	yes	yes	fixed merge
CRPD	yes	yes	shallow tree ensemble

Table 4 Test effectiveness for single retrievers, fixed hybrid rules, the training-free adaptive fusion baseline, and CRPD. Recall@10 and Recall@100 are abbreviated R@10 and R@100.

Method	SciFact nDCG@10	R@10	R@100	ANTIQUE nDCG@10	R@10	R@100
BM25	0.6567	0.7726	0.8726	0.2398	0.1888	0.4422
BGE-M3	0.6417	0.7751	0.9037	0.3019	0.2440	0.5223
RRF	0.6897	0.8154	0.9477	0.3578	0.2843	**0.6075**
Fusion-0.25	0.6950	0.8081	0.9199	0.3028	0.2416	0.5196
Fusion-0.50	0.7101	0.8346	0.9443	**0.3597**	**0.2904**	0.6044
Fusion-0.75	0.6938	0.8184	0.9210	0.3440	0.2657	0.5778
Dynamic fusion	0.7099	0.8318	0.9443	0.3430	0.2705	0.5809
CRPD	**0.7174**	**0.8418**	**0.9510**	0.3592	0.2852	0.5995

Table 5 Comparison with independently learned adaptive alternatives on the locked test splits. Frozen BGE-M3 query embeddings and the prespecified composite utility are the adaptation targets of QDAP-S. Bold marks the best value within this table.

Method	SciFact nDCG@10	R@10	R@100	ANTIQUE nDCG@10	R@10	R@100
Policy classifier (soft)	0.7022	0.8234	0.9410	0.3435	0.2669	**0.6003**
QDAP-S (adapted)	0.7083	0.8389	0.9443	0.3405	0.2619	0.5931
CRPD	**0.7174**	**0.8418**	**0.9510**	**0.3592**	**0.2852**	0.5995

SciFact and 0.3416 on ANTIQUE, numbers that climb to 0.7174 and 0.3592 under the soft policy. Competitive though it remains, the query-only student sits below the full CRPD model on SciFact, a pattern consistent with some additional contribution from retrieval-state information beyond the query.

Table 6 Mechanism and target-parameterization controls. A development-selected optional fallback stayed inactive on both datasets and is omitted as numerically identical to CRPD. Bold marks the best value within this table.

Method	SciFact nDCG@10	R@10	R@100	ANTIQUE nDCG@10	R@10	R@100
CRPD (hard decoding)	0.7055	0.8318	0.9477	0.3416	0.2644	0.5775
CRPD (query-only)	0.7078	0.8362	**0.9510**	0.3602	0.2843	0.5997
Fixed portfolio mixture	0.7061	0.8329	0.9443	**0.3606**	**0.2864**	**0.6046**
Direct utility regression	**0.7210**	**0.8462**	**0.9510**	0.3575	0.2787	0.6044
CRPD	0.7174	0.8418	**0.9510**	0.3592	0.2852	0.5995

5.5 External low-heterogeneity boundary

ANTIQUE provides a test case for querying heterogeneity with limited variability, in which the benefits brought by adaptation may be relatively small. Under the offline criterion, for 92.0% of the queries, a single fusion strategy performed the best. The query-independent combination mixing strategy achieved the highest nDCG@10 value of 0.3606, which was only 0.0014 higher than CRPD. The confidence interval for this pairing includes the zero value. Meanwhile, CRPD significantly outperformed the independently trained strategy classifier and QDAP-S. These results indicate that when a single global strategy has performed well on most queries, the space for query adaptation is limited. Nevertheless, CRPD remains competitive with other learning-based adaptive methods.

5.6 Repeated-partition stability

We re-fitted the student model on five fixed subsets, each containing 80% of the official SciFact training queries. All five runs maintained the performance advantage. The average nDCG@10 reached 0.7158±0.0021, even the lowest result (0.7135) was higher than RRF (0.6897) and the strongest fixed fusion strategy (0.7101). In all five runs, Recall@100 remained at 0.9510. Consistent results indicate that the improvement effect of CRPD on SciFact remains stable across different training shards.

5.7 Adaptation-opportunity analysis

Figure 4 analyzes how the adaptive benefits change with the variations in claim characteristics and search conditions. When the overlap degree of BM25 and the dense results is low, the benefit of CRPD is the greatest. In the lowest overlap quartile, CRPD increases nDCG@10 compared to RRF by 0.0585, while in the highest overlap quartile, the increase is only 0.0088. For queries with low Oracle strategy entropy, this benefit can reach 0.0935, but for high entropy quartiles, no improvement is observed. These results indicate that when the two retrievers have differences and one strategy has a significant practical advantage, the adaptive effect is the most significant.

5.8 Claim-level evidence cases

The first two cases illustrate different weighted strategies. For claim 1241, CRPD allocates most of the weight to dense retrieval and dense-high fusion strategies. For claim 768, the weights are distributed among RRF, balanced fusion, and BM25. In neutral conditions, both systems did not include relevant evidence in the top 10. In failure cases, the student over-assigned weight to the dense-high strategy, resulting in the judged evidence ranking lower than its position under fixed fusion. This case indicates that query adaptive weighting may also lead to poor ranking results.

5.9 Controller overhead

On SciFact, the total time consumed for feature generation and median policy prediction for each query is 0.961 milliseconds. In contrast, on the same machine, the median latency of dense retrieval by BGE-M3 is 13.25 milliseconds per query. Therefore, CRPD adds a slight computational overhead after retrieval, while the retriever itself remains unchanged.

6 Discussion

6.1 What is learned by the policy?

Across the action-specific students, the most important features describe how scores are distributed in

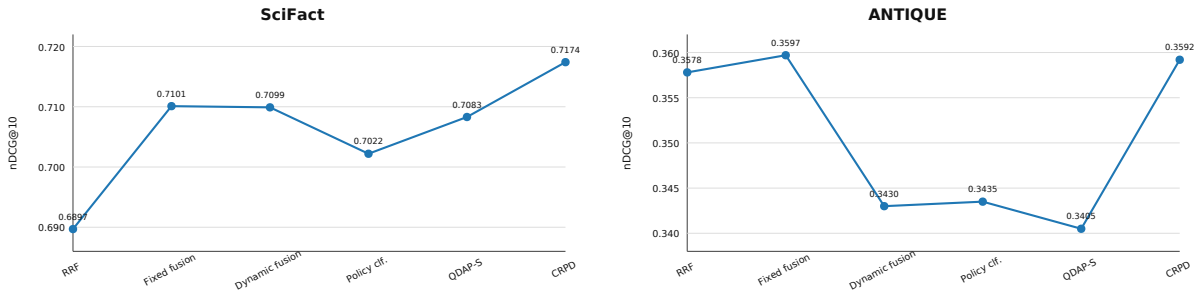

Fig. 2 Test nDCG@10 across representative fixed/training-free methods and independently learned adaptive alternatives. CRPD has the highest point estimate among the learned adaptive methods on both datasets; ANTIQUE retains a slightly stronger fixed fusion.

Table 7 Heterogeneity of the per-query offline oracle. Active-strategy count and dominant-action share summarize the offline action landscape.

Dataset	Queries	Active actions	Dominant action	Share
SCIFACT	300	6	fusion_075	45.3%
ANTIQUE	200	4	fusion_075	92.0%

Table 8 SciFact stability across five fixed training partitions.

	Seed	nDCG@10	Recall@10	Recall@100
	20260728	0.7185	0.8418	0.9510
	20260729	0.7168	0.8418	0.9510
	20260730	0.7165	0.8418	0.9510
	20260731	0.7135	0.8412	0.9510
	20260732	0.7139	0.8429	0.9510
Average ± Standard Deviation		0.7158±0.0021	0.8419	0.9510

Table 9 Selected SciFact query scenarios. Gain is CRPD minus RRF nDCG@10.

Query Group	Queries	CRPD nDCG@10	Compared to RRF's improvement
Length / Long	100	0.7487	+0.0259
Length / Short	100	0.7209	+0.0367
Jaccard / Q1	75	0.6595	+0.0585
Jaccard / Q4	75	0.6803	+0.0088
Information Entropy / Q1	75	0.5487	+0.0935
Information Entropy / Q4	75	0.8635	+0.0000

the retrieval results, particularly their concentration and variation. Query length has much lower importance. Although the model also uses overlap and rank disagreement features, neither appears among the top five. Therefore, we believe that low overlap observed in the scenario analysis can serve as a useful signal for identifying opportunities for query adaptation, but it cannot be considered the sole cause of performance improvement.

Figure 3 shows why CRPD performs differently on the two datasets. On SciFact, all six strategies appear as the best action for some queries, and the most frequent strategy accounts for only 45.3% of queries. On ANTIQUE, Fusion-0.75 has

Table 10 Real SciFact cases. Rank 1001 denotes absence from the saved Top-1000 run.

Case	Scientific claim	Evidence ID	Fusion rank	CRPD rank	ΔnDCG
improve_1	The myocardial lineage develops from cardiac progenitors of mesodermal origin.	4427392	8	1	+0.685
improve_2	Mercaptopurine is anabolized into the inactive methylmercaptopurine by thiopurine methyltrasnferase (TPMT).	6421792	50	4	+0.431
neutral	0-dimensional biomaterials show inductive properties.	31715818	227	229	+0.000
failure	Up-regulation of the p53 pathway and related molecular events casues cancer resistance and results in a significantly shortened lifespan marked by senescent cells and accelerated organismal aging.	9559146	1	2	-0.369

Table 11 Measured CRPD control-layer footprint on SciFact.

Component	Measurement Method
Feature Generation	Each query takes 0.518 milliseconds
Policy Prediction p50	Each query takes 0.443 milliseconds
Policy Prediction p95	Each query takes 0.457 milliseconds
Serialized ensemble	3.26 MiB

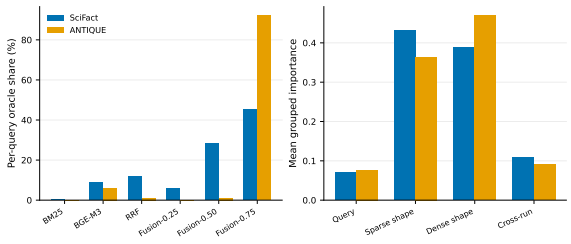

Fig. 3 Offline action heterogeneity (left) and grouped student feature importance (right). SciFact distributes oracle choices across the portfolio, whereas ANTIQUE is dominated by a single fixed-fusion action. Values come from saved oracle counts and training feature-importance artifacts.

a large share in the oracle distribution, reaching 92.0%. This means that most queries favor the same strategy. If the best strategy varies greatly across queries, CRPD has more room to improve. When one strategy works best for nearly all queries, a fixed strategy can already achieve strong results. Table 7 presents this difference. ANTIQUE therefore shows that query adaptation offers limited benefits when one strategy dominates.

6.2 Operational interpretation

CRPD can be added to a RAG system that already uses both sparse and dense retrieval. We first build the strategy–utility matrix offline using historical queries with relevance judgments. We then deploy the trained student as a lightweight controller, stored in a 3.26 MiB model file (Table 2). During inference, the system produces the same sparse and dense retrieval results as before. The controller replaces fixed fusion with regret-based adaptive fusion by reading both retrieval outputs and predicting regret for each strategy. Other parts of the retrieval pipeline remain unchanged. Table 2 shows that the additional computation is small compared with the cost of dense retrieval.

6.3 What the learned comparisons show

The comparisons with learned baselines show two main results. First, CRPD achieves higher scores than the adapted QDAP-S and the independently trained policy classifier on both datasets. The differences are statistically significant on ANTIQUE, but not on SciFact. This suggests that learning the relative performance of all candidate strategies may help improve retrieval. The policy classifier learns a single best-action label, while QDAP-S predicts a sparse–dense mixing coefficient. CRPD also performs worse when hard decoding replaces soft fusion. Together, these results support the use of continuous multi-action supervision and soft strategy fusion.

Second, regret regression and direct utility regression achieve similar performance. On

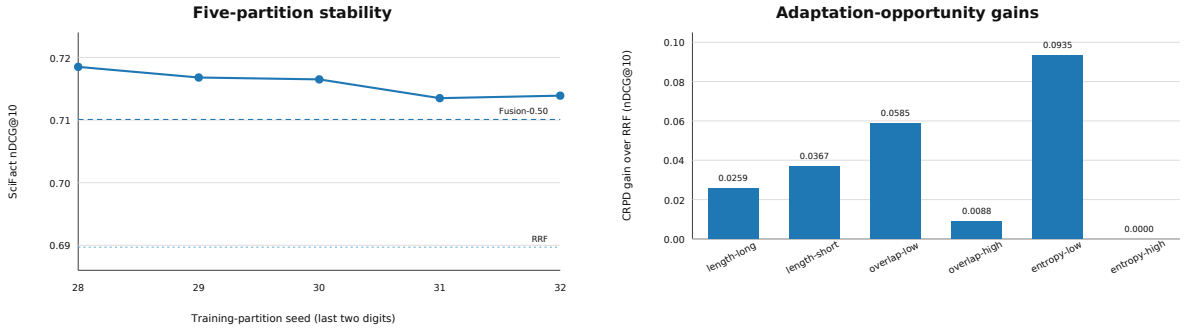

Fig. 4 SciFact stability and adaptation-opportunity analysis. Every refit remains above the fixed references, and the largest gains occur under low-overlap or otherwise high-opportunity retrieval conditions.

SciFact, the capacity-matched utility regressor reaches 0.7210 nDCG@10, compared with 0.7174 for CRPD. CRPD performs slightly better on ANTIQUE. Neither difference is statistically significant. This result is consistent with the identity $\text{softmax}(-\Delta/\tau) = \text{softmax}(U/\tau)$. The experiments suggest that continuous multi-action supervision and soft fusion contribute more to performance than the choice between regret and utility targets. Regret is still useful because it is nonnegative and measures the utility loss relative to the best strategy for each query. This makes it useful for training analysis and deployment monitoring.

6.4 Failure modes and decision rules

CRPD may fail under three conditions. First, one strategy may perform best for almost all queries, leaving little room for query adaptation. Second, the available features may not predict which strategies work best for a given query. Third, the student may predict reasonable strategy weights but still produce a poor ranking because the scores are not properly balanced during fusion. We use hard decoding, query-only features, and soft fusion ablations to examine these problems.

ANTIQUE shows the first failure case. Its training oracle favors one strategy for most queries, and the query-independent portfolio mixture achieves performance close to the learned policy. SciFact has a more diverse oracle distribution, with all six strategies selected across different queries. On SciFact, CRPD outperforms RRF and achieves a higher nDCG@10 point estimate than the strongest fixed fusion strategy. When strategy utilities differ across queries and these differences are predictable, adaptive control can improve retrieval performance.

6.5 Limitations

Our evaluation covers two datasets and one fixed pair of sparse and dense retrievers. The adapted QDAP-S uses frozen BGE-M3 query embeddings and our composite utility target, so its configuration differs from the original QDAP system. The utility weights $(w_1, w_2, w_3) = (0.5, 0.2, 0.3)$ give greater importance to early ranking quality. Other applications may require different weights. The oracle represents the best available strategy within our six-action portfolio, and our evaluation focuses on retrieval quality without measuring downstream answer generation. The capacity-matched experiments also show no significant difference between regret and direct utility regression. On ANTIQUE, a strong fixed mixture matches the learned policy because one strategy dominates most queries. Future work should test CRPD on more scientific domains and retrieval settings.

7 Conclusion

From a frozen sparse–dense portfolio, CRPD builds a query-adaptive ranker for scientific evidence, with full-information supervision offline and a small regret-predicting controller online. SciFact yields 0.7174 nDCG@10, a +0.0276 improvement over RRF (Holm-adjusted $p = 0.0016$), together with 15.4% closure of the static-to-oracle utility gap. Both the adapted QDAP-S and the independently trained policy classifier finish below CRPD on the

Table 12 Deployment decision guide derived from the observed method boundaries.

Observed condition	Recommended policy	Reason
One action dominates	Fixed fusion	Little policy heterogeneity
Actions vary, features predict regret	CRPD	Query adaptation is learnable
No judged development queries	RRF or fixed fusion	Regret teacher cannot be fitted
Only one retrieval run is available	Single retriever	CRPD does not remove retrieval cost

two datasets, significantly so on ANTIQUE. Across all five SciFact training partitions the strongest prespecified fixed fusion is cleared, and in the lowest sparse–dense overlap quartile, exactly where adaptation opportunity concentrates, the RRF gain widens to +0.0585. Landing in the same effectiveness region, the capacity-matched direct-utility student supports a claim of parity between the two continuous target forms rather than superiority of regret centering. Measured feature generation and median prediction add 0.961 ms per query, no retriever undergoes retraining, and no language model is called anywhere in the loop. Where one portfolio action dominates, as ANTIQUE shows in drawing the deployment boundary, the better choice is a strong fixed mixture. Natural next steps cover further scientific domains, an explicit retrieval-cost term inside the teacher utility, and drift monitoring as judged queries accumulate.

Acknowledgements. Not applicable.

Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest. The authors declare that they have no competing interest.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. Data sets generated during the current study are available from the corresponding author on reasonable request.

Author contributions. Xiao Liu: Conceptualization, Methodology, Investigation, Writing – Original Draft. Yuhan Dai: Software, Validation, Data Curation. Wenjie Chen: Validation, Investigation, Data Curation. Yaomu Tan: Formal Analysis, Visualization. Zhenjie Han: Methodology, Writing – Review & Editing. Ziang Yao: Conceptualization, Supervision, Project Administration, Writing – Review & Editing. All authors read and approved the final manuscript.

References

[1] Arabzadeh N, Yan X, Clarke CLA. Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection. arXiv preprint arXiv:210910739. 2021;https://doi.org/10.48550/arXiv.2109.10739.

[2] Kim J, Yoon S, Le XB, Nam Y, Kim D, Song H, et al. QuDAR: Query-Wise Dual-Perspective Adaptive Retrieval. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics; 2026. p. 38662–38679.

[3] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval. 2009;4(1–2):1–174. https://doi.org/10.1561/1500000019.

[4] Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2020. p. 6769–6781.

[5] Formal T, Piwowarski B, Clinchant S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In: Proceedings of the 44th International ACM SIGIR

Conference on Research and Development in Information Retrieval. ACM; 2021. p. 2288–2292.

[6] Khattab O, Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2020. p. 39–48.

[7] Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:240203216. 2024;https://doi.org/10.48550/arXiv.2402.03216.

[8] Cormack GV, Clarke CLA, Büttcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2009. p. 758–759.

[9] Bruch S, Gai S, Ingber A. An Analysis of Fusion Functions for Hybrid Retrieval. ACM Transactions on Information Systems. 2023;42(1):1–35. https://doi.org/10.1145/3596512.

[10] Lee D, Hwang Sw, Lee K, Choi S, Park S. On Complementarity Objectives for Hybrid Retrieval. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023. p. 13357–13368.

[11] Posokhov P, Skrylnikov S, Masliukhin S, Zavgorodniaia A, Koroteeva O, Matveev Y. Query-Adaptive Hybrid Search. Machine Learning and Knowledge Extraction. 2026;8(4):91. https://doi.org/10.3390/make8040091.

[12] Yuan A, Zhang H, Nazarian S. AgentIR: A Workload-Adaptive Cascade Retrieval Substrate for Long-Term Conversational Memory. arXiv preprint arXiv:260525092. 2026;https://doi.org/10.48550/arXiv.2605.25092.

[13] Genc AC, Korukluoglu EK, Allan J. Adaptive Re-Ranking. arXiv preprint arXiv:260625249. 2026;https://doi.org/10.48550/arXiv.2606.25249.

[14] Pezo I, Hanbury A, Staudinger M. ipezoTU at SemEval-2025 Task 7: Hybrid Ensemble Retrieval for Multilingual Fact-Checking. In: Proceedings of the 19th International Workshop on Semantic Evaluation. Association for Computational Linguistics; 2025. p. 1159–1167. Available from: https://aclanthology.org/2025.semeval-1.153/.

[15] Kiepura A, Lam J. ClimateCheck2025: Multi-Stage Retrieval Meets LLMs for Automated Scientific Fact-Checking. In: Proceedings of the Fourth Workshop on Scholarly Document Processing. Association for Computational Linguistics; 2025. p. 293–306.

[16] Tang X, Gao Q, Li J, Du N, Li Q, Xie S. MBA-RAG: A Bandit Approach for Adaptive Retrieval-Augmented Generation through Question Complexity. In: Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics; 2025. p. 3248–3254. Available from: https://aclanthology.org/2025.coling-main.218/.

[17] Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020. p. 7534–7550.

[18] Hashemi H, Aliannejadi M, Zamani H, Croft WB. ANTIQUE: A Non-Factoid Question Answering Benchmark. In: Advances in Information Retrieval. vol. 12036 of Lecture Notes in Computer Science; 2020. p. 166–173.

[19] Lin J, Ma X, Lin SC, Yang JH, Pradeep R, Nogueira R. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2021. p. 2356–2362.