RETRIEVER ADAPTERS AS TEACHERS: REPRESENTATION CORRECTION DISTILLATION FOR DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting dense retrievers to new target domains is severely constrained by the high cost of manual relevance labeling. While lightweight adapters can learn domain specialization from limited labeled data, they leave abundant unlabeled queries unused. To address this, we introduce Representation Correction Distillation (RCD), a semi-supervised framework that transfers adapter-learned specialization into a standalone dense retriever using both labeled and additional unlabeled target queries, without using relevance labels during student training. Rather than matching full representation spaces, RCD explicitly isolates and distills the task-specific representation updates induced by the adapter, preserving the base model's pre-existing structure while integrating new domain preferences via ranking distillation. Consequently, the distillation phase requires zero additional relevance labels. Across five benchmark datasets, RCD improves macro-average nDCG over the base retriever by up to 21.7%. When the teacher is trained on only 50% of the labeled queries, RCD uses additional unlabeled target queries to train a student that outperforms the teacher by 3.7% in macro-average nDCG.

1 INTRODUCTION

Dense retrievers are a core component of modern search and retrieval-augmented generation (RAG) systems (Lewis et al., 2020), yet retrievers pretrained on broad corpora are not necessarily aligned with what constitutes relevance in a particular target domain. Their effectiveness can deteriorate under domain and task shifts, especially in specialized and continually evolving settings (Chen et al., 2022; Ko et al., 2025; Thakur et al., 2021). Such failures have been observed in domains including law, finance, biomedical, and scientific literature, where retrieval depends on specialized terminology and task-specific information needs (Thakur et al., 2021; Song et al., 2025). Dense retrievers also struggle with compositionally constrained queries involving conjunction, disjunction, negation, and multi-hop relations, where semantic similarity alone may not capture all relevance conditions (Xu et al., 2026; Petcu et al., 2025). Adapting the retriever to these settings can improve task-specific relevance, but doing so typically requires query-document relevance labels that are expensive to obtain at scale (Karpukhin et al., 2020; Hofstätter et al., 2021). In a legal or biomedical collection, each label requires a domain expert to read a query and judge its candidate passages. This creates a ***domain-adaptation gap***, in which relevance labels are scarce while unlabeled target-domain data is abundant (Wang et al., 2022a).

Direct supervised fine-tuning updates the retriever using available query-document relevance labels and can provide strong task-specific adaptation (Karpukhin et al., 2020), but its effectiveness is necessarily tied to the amount of labeled data available. Qrel-free approaches or methods that do not require manually annotated relevance labels instead exploit unlabeled target-domain data through self-supervised objectives, synthetic queries, pseudo-labels, or teacher-generated signals (Wang et al., 2022a; Saad-Falcon et al., 2023), avoiding additional manual labeling but relying on indirect supervision. A complementary line of work uses lightweight retriever adapters, including Search-Adaptor (Yoon et al., 2024a), IMRNNs (Interpretable Modular Retrieval Neural Networks) (Saxena et al., 2026), and related parameter-efficient adaptation methods (Maekawa et al., 2026; Sanjeev & Troynikov, 2024; Zeighami et al., 2025). These methods keep the pretrained retriever largely or entirely frozen and learn a small task-specific transformation using limited relevance labels.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

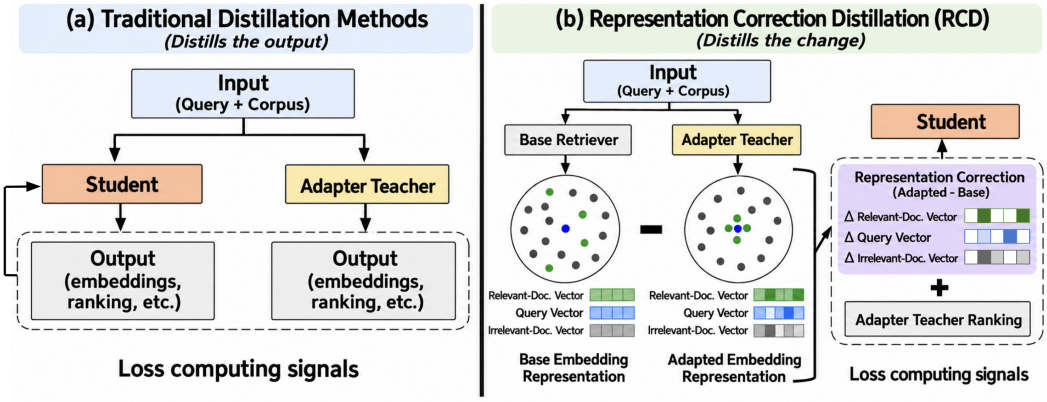


Figure 1: Overview of Representation Correction Distillation (RCD). (a) Traditional distillation transfers a teacher's final outputs, such as embeddings or ranking scores, to the student. (b) RCD instead distills the task-specific representation changes induced by a retriever adapter relative to the frozen base retriever, together with the adapter's ranking behavior. The resulting student directly incorporates the adapter-learned specialization and is used as a standalone retriever at inference.

While adapters are attractive when relevance labels are scarce, their supervised training is limited to the available labeled data. Once these labels have been used, additional unlabeled target-domain queries cannot provide further supervision under the same training objective. The resulting adapter therefore captures task-specific specialization from scarce labels, while a larger pool of unlabeled target data may remain unused for further propagating that specialization.

This observation suggests a broader role for retriever adapters, where specialization learned from scarce relevance labels can extend supervision to target queries unseen during adapter training. We therefore ask: *can an adapter trained with limited labels be used to further improve its underlying retriever using unlabeled target-domain queries?* Taking inspiration from Rocchio's relevance-feedback method (Rocchio Jr, 1971), which adds a correction vector to a query to move it toward relevant documents. An adapter can be viewed the same way, as a learned correction added to each base embedding. We therefore treat the representation changes learned by a task-specific adapter as a transferable supervision signal. We introduce Representation Correction Distillation (RCD), a semi-supervised *adapter-to-retriever* framework in which a trained and frozen adapter serves as a task-specific teacher. Because both the adapted retriever and the student are initialized from the same pretrained retriever, they begin with the same representation structure. The new information introduced by adaptation therefore lies in how those representations are altered.

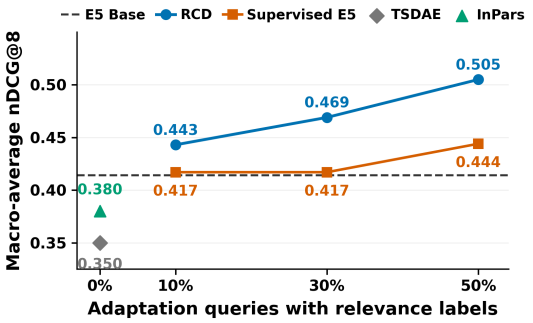


Figure 2: Macro-average nDCG@8 across five datasets at different relevance-supervision levels. The dashed line shows the E5 baseline.

RCD therefore isolates these adapter-induced changes and aligns their directions rather than requiring the student to reproduce the complete adapted representation or its exact displacement magnitude. Figure 1 contrasts RCD with conventional distillation methods that transfer final scores or representations. These signals combined allow the student to propagate the adapter's learned specialization over a broader pool of target queries without using relevance labels during student optimization, while producing a standalone dense retriever at inference time.

We evaluate RCD on five retrieval benchmarks spanning domain-specific and compositionally constrained retrieval, using `intfloat/e5-large-v2` (Wang et al., 2022b) as the base retriever and IMRNNs as the adapter teacher. RCD improves the base retriever by up to 21.7% in nDCG@8

and, on average, surpasses its own adapter teacher at every supervision budget (Figure 2). This improvement over the teacher is attributable to the additional unlabeled queries, since a student trained only on the labeled queries used to train the teacher performs on par with it rather than exceeding it. The transfer further holds with a second adapter architecture, Search-Adaptor (Yoon et al., 2024a), and a second retriever backbone, Qwen3-Embedding (Zhang et al., 2025).

Our contributions are threefold:

- We reframe retriever adapters as teachers rather than endpoints of adaptation. A trained adapter can improve its own base retriever using unlabeled target queries, and the resulting retriever needs no adapter at inference.
- We identify what should be distilled when teacher and student share a base model. The specialization lies in the *representation change* rather than the *final representation*, and distilling the change transfers it while distilling the full representation degrades retrieval.
- We show that unlabeled target queries, not additional labels, allow the student to exceed its teacher. This makes cheap, abundant queries a direct source of retrieval improvement.

2 Related Work

Dense Retriever Adaptation. Adapting retrieval models to task- or domain-specific notions of relevance has a long history, from classical relevance-feedback methods such as Rocchio Jr (1971) to modern dense retrievers such as DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2020), which learn query and document representations from supervised query-passage pairs. However, dense retrievers trained on one distribution can generalize unevenly to unseen domains, as highlighted by heterogeneous evaluations such as BEIR (Thakur et al., 2021). This has motivated approaches that reduce reliance on target-domain relevance labels. TSDAE (Wang et al., 2021) and Contriever (Izacard et al., 2022) learn from unlabeled text through self-supervised objectives, while MoDIR (Xin et al., 2022) and COCO-DR (Yu et al., 2022) explicitly address domain shift through domain-invariant or target-corpus representation learning. Although these methods enable adaptation without manually annotated target-domain relevance labels, their gains are marginal and are not robust in performance across retrieval domains. A complementary line of work constructs task-specific supervision automatically: GPL (Wang et al., 2022a) generates synthetic queries and cross-encoder pseudo-labels, while InPars (Bonifacio et al., 2022) uses language models to generate synthetic query-document training examples. Promptagator (Dai et al., 2022) uses a small number of examples to generate larger task-specific training sets, and more recent approaches such as DUQGen (Chandradevan et al., 2024) and tRAG (Lee et al., 2025) further improve synthetic target-domain supervision. While effective, these approaches typically require additional generation, scoring, or pseudo-labeling stages, increasing adaptation-time computation and pipeline complexity.

Retriever Adapters. Parameter-efficient adapters were originally introduced as a way to specialize pretrained models while keeping most of their parameters fixed (Houlsby et al., 2019), and have since been explored for dense representation learning and information retrieval. Early work applied adapters to dense retrievers for parameter-efficient cross-lingual transfer (Yang et al., 2022), while more recent methods operate directly on frozen retrieval representations. Search-Adaptor learns a lightweight transformation that customizes query and document embeddings for a target retrieval task (Yoon et al., 2024a), and Matryoshka-Adaptor extends this idea toward adapting embeddings at reduced dimensionalities (Yoon et al., 2024b). IMRNNs instead learn query- and document-dependent embedding modulations, allowing task-specific changes to the retrieval space while retaining the underlying encoder (Saxena et al., 2026). Recent work such as *Align then Train* further studies label-efficient retriever adapters by combining self-supervised alignment with limited relevance supervision (Maekawa et al., 2026). Related post-hoc approaches such as NUDGE similarly improve frozen embedding spaces without conventional end-to-end fine-tuning (Zeighami et al., 2025). Collectively, these methods show that lightweight transformations around a frozen retriever can capture substantial task-specific specialization. However, they treat the learned adapter or transformed embedding space as the endpoint of adaptation; they do not study whether the specialization encoded by a trained retriever adapter can subsequently supervise updates to the underlying retriever on additional unlabeled target queries.

Knowledge Distillation for Dense Retrieval. Knowledge distillation has been widely used to transfer retrieval behavior from stronger teachers into dense retrievers. Score-based approaches

such as Margin-MSE supervise the student using differences in teacher relevance scores (Hofstätter et al., 2020), while methods such as RocketQAv2 (Ren et al., 2021), PROD (Lin et al., 2023), and PairDistill (Huang & Chen, 2024) transfer richer ranking information. Readers and rerankers have also been used as intermediate teachers for retrievers (Izacard & Grave, 2020; Saad-Falcon et al., 2023). Beyond ranking scores, EmbedDistill provides representation-level supervision by matching the teacher's embedding space (Kim et al., 2023). AdapterDistillation similarly uses trained adapters as knowledge sources, but distills their representations into another adapter while keeping the underlying pretrained model frozen (Wang et al., 2023). Prior feature distillation work has also shown that angular or normalized alignment can be more effective than requiring exact feature-magnitude matching (Liu et al., 2024). RCD differs from these approaches in the object being distilled. The adapter and student inherit substantial representation structure from the same pretrained retriever, while adaptation introduces a task-specific change to that structure. Rather than supervising the complete adapted representation, RCD isolates this change and transfers its direction together with the adapter's ranking behavior. This allows the representation objective to focus on the specialization introduced by adaptation instead of reproducing information already present in the base retriever.

3 METHOD

3.1 PROBLEM SETUP

Let $\mathcal{C}$ denote a target-domain document corpus and f_0 a pretrained dense retriever. We assume access to a limited set of labeled target queries $\mathcal{Q}_L$, for which query-document relevance labels are available, and an additional set of unlabeled target queries $\mathcal{Q}_U$.

A lightweight retriever adapter g_ψ is first trained using the relevance labels available for $\mathcal{Q}_L$, while the base retriever f_0 remains frozen. After training, the adapter is frozen and used as a teacher for a student retriever f_θ initialized from f_0.

The student is trained over target queries from $\mathcal{Q} = \mathcal{Q}_L \cup \mathcal{Q}_U$, without using ground-truth relevance labels during student training.

For a query q and document d, the base retriever produces $\mathbf{q}_0 = f_0(q)$ and $\mathbf{d}_0 = f_0(d)$, with retrieval score $s_0(q, d) = \cos(\mathbf{q}_0, \mathbf{d}_0)$. For each training query, we retrieve a candidate set $\mathcal{D}_q = \mathrm{TopK}_{d \in \mathcal{C}}\, s_0(q, d)$ using the frozen base retriever.

The frozen adapter then provides task-adapted representations and retrieval scores for these candidates. RCD uses these signals to transfer the specialization learned by the adapter into the student retriever.

3.2 REPRESENTATION CORRECTION DISTILLATION

Representation correction. The main idea of RCD is to distill the *change* the adapter introduces rather than directly matching its final representations. Let $\mathbf{q}_T$ and $\mathbf{d}_T$ denote the query and document representations produced by the frozen adapter. We define the adapter-induced representation corrections as

$$\Delta\mathbf{q}_T = \mathbf{q}_T - \mathbf{q}_0, \qquad \Delta\mathbf{d}_T = \mathbf{d}_T - \mathbf{d}_0. \tag{1}$$

Similarly, for the student retriever,

$$\Delta\mathbf{q}_\theta = f_\theta(q) - \mathbf{q}_0, \qquad \Delta\mathbf{d}_\theta = f_\theta(d) - \mathbf{d}_0. \tag{2}$$

RCD uses cosine similarity because the correction direction, rather than its exact magnitude, is the representation signal we seek to transfer. Raw Euclidean matching constrains both quantities, requiring the standalone retriever to reproduce the adapter's displacement norm in addition to its direction. Prior knowledge-distillation work has similarly found angular or scale-invariant feature alignment useful when exact feature magnitude need not be preserved (Liu et al., 2024). We therefore use cosine alignment for the correction signal and use ranking distillation to constrain the retrieval behavior produced by the resulting representations.

This alignment is applied to the corrections rather than the final representations. From Eqs. (1) and (2), both the teacher and student representations contain the shared base component $\mathbf{q}_0$. When the corrections are small relative to this shared representation, the final representations can remain highly similar even when the correction directions disagree. Aligning $\Delta\mathbf{q}_\theta$ with $\Delta\mathbf{q}_T$ instead removes this

shared component and directly measures the direction of the adaptation-induced change. The same argument applies to $\Delta\mathbf{d}_\theta$ and $\Delta\mathbf{d}_T$. Additionally, since $\Delta\mathbf{q}_\theta = \mathbf{0}$ at initialization, we add a small constant ϵ to the correction norms to keep the cosine well-defined

We define the representation-correction loss as

$$\mathcal{L}_{\text{corr}}(q) = \frac{1}{2}\left[1 - \cos\left(\Delta\mathbf{q}_\theta, \Delta\mathbf{q}_T\right) + \frac{1}{|\mathcal{D}_q|}\sum_{d\in\mathcal{D}_q}\left(1 - \cos\left(\Delta\mathbf{d}_\theta, \Delta\mathbf{d}_T\right)\right)\right]. \tag{3}$$

This objective explicitly transfers how the adapter changes the original retrieval space, rather than supervising the student only with the adapter's final representations.

Ranking distillation. In addition to representation correction, we use the adapter's ranking behavior as an additional supervision signal. For each candidate $d \in \mathcal{D}_q$, the adapter produces a score $s_T(q, d)$, which is converted into a probability distribution:

$$p_T(d \mid q) = \frac{\exp(s_T(q, d)/\tau)}{\sum_{d'\in\mathcal{D}_q}\exp(s_T(q, d')/\tau)}, \tag{4}$$

where τ is the distillation temperature. The student produces scores $s_\theta(q, d) = \cos(f_\theta(q), f_\theta(d))$, with $p_\theta(d \mid q)$ defined analogously. We distill the adapter's ranking behavior using $\mathcal{L}_{\text{rank}}(q) = D_{\text{KL}}\big(p_T(\cdot \mid q) \,\|\, p_\theta(\cdot \mid q)\big)$.

Training objective. The RCD objective combines representation correction with auxiliary ranking distillation:

$$\mathcal{L}_{\text{RCD}} = \mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{rank}}. \tag{5}$$

No relevance labels are used during student optimization. After training, the adapter is no longer required and retrieval is performed directly using the student retriever f_θ.

4 Experimental Setup

Table 1: **Dataset statistics and relevance-supervision budgets.** $|C|$ denotes the number of documents in the corpus, and $|Q_{\text{adapt}}|$ denotes the training-query pool for adaptation. Columns: 10%, 30%, and 50% report the number of queries with relevance labels used to train the adapter teacher.

Dataset	Domain / Task	Hop Type	$\lvert C\rvert$	$\lvert Q_{\text{adapt}}\rvert$	10%	30%	50%
ContractNLI	Legal / Contracts	Multi-hop	95	783	78	235	392
PrivacyQA	Privacy	Multi-hop	7	156	16	47	78
ComLQ	Complex logical queries	Mixed	11,251	2,329	233	699	1,165
SciFact	Scientific evidence	Single-hop	5,183	728	73	219	364
NFCorpus	Biomedical	Single-hop	3,633	2,590	259	777	1,295

Datasets and Evaluation. We evaluate RCD on five retrieval benchmarks: ContractNLI, PrivacyQA (Pipitone & Alami, 2024), ComLQ (Xu et al., 2026), SciFact, and NFCorpus (Thakur et al., 2021). The suite spans heterogeneous domains and includes both single-hop and multi-hop retrieval settings. ComLQ additionally provides substantial within-benchmark diversity, covering multiple reasoning-hop depths and different logical query structures, including conjunction, negation, and projection. Table 1 summarizes dataset statistics. For each dataset, we split queries into training, validation, and test sets: 80% for training, 10% for validation, and 10% for testing. The validation and test queries are held out from all adaptation and student training. We refer to the training split as the adaptation pool. To study adaptation under limited relevance supervision, we consider budgets of 10%, 30%, and 50%, corresponding to the fraction of adaptation-pool queries whose relevance labels are used to train the adapter and teacher. The remaining queries in the adaptation pool are treated as unlabeled: their query texts may subsequently be used during RCD training, but their relevance labels are never exposed to the student. We use nDCG@8 (Järvelin & Kekäläinen, 2002) as the primary evaluation metric and additionally report nDCG at $k \in \{2, 4, 8, 16, 32, 64\}$ in the appendix B. For all learned methods, we report the mean performance over three optimization seeds (42, 43, and 44).

Supervision Protocol. For each dataset, we partition the adaptation queries into a labeled subset Q_L and an unlabeled subset Q_U. The relevance labels of Q_L are used only to train the adapter teacher. The teacher is then frozen, and a student initialized from the same base retriever is distilled from it without relevance labels. To isolate the contribution of unlabeled queries, we compare two student variants that share the same frozen teacher and optimization budget and differ only in their training queries: RCD-ALLQUERIES uses $Q_L \cup Q_U$, whereas RCD-LABELEDONLY uses Q_L alone. Validation and test queries are excluded from all adapter and student training.

Base Retriever and Adapter Teachers. Our primary experiments use `intfloat/e5-large-v2` as the base retriever and IMRNNs (Saxena et al., 2026) as the adapter teacher, with a separate teacher trained for each supervision budget while the E5 encoder remains frozen. To test whether RCD generalizes beyond this configuration, we replace IMRNNs with Search-Adaptor (Yoon et al., 2024a) to vary the adapter architecture, and replace E5 with Qwen3-Embedding (Zhang et al., 2025) to vary the retriever backbone.

Baselines. We compare RCD against the pretrained base retriever, same-budget supervised fine-tuning, the frozen adapter teacher, conventional distillation methods, and target-domain adaptation baselines. SUPERVISED E5 fine-tunes the base retriever using the same relevance-supervision budget used to train the adapter teacher. We include TSDAE (Wang et al., 2021) as an unsupervised target-domain adaptation baseline and InPars (Bonifacio et al., 2022) as a synthetic-supervision baseline. For distillation, Margin-MSE (Hofstätter et al., 2020) transfers teacher score differences; because student training uses no relevance labels, we pair the teacher's highest-scoring candidate with the remaining candidates. EmbedDistill (Kim et al., 2023) distills the teacher's final query and document representations together with its ranking scores. We also report the frozen adapter teacher to measure whether RCD can improve beyond the specialization learned by the adapter itself.

Training Setup. For each training query, we retrieve the top-64 candidate documents using the frozen E5 base retriever. We set the candidate pool size to 64 to match the largest retrieval cutoff evaluated in our experiments of $k = 64$ (refer to appendix B for results across different k values). Candidate construction does not use relevance labels or inject known relevant documents. RCD, Margin-MSE, and EmbedDistill use the same candidate sets and frozen adapter outputs, and all students are initialized from the same E5 checkpoint. Student optimization uses no relevance labels. We keep the training configuration and optimization budget fixed across distillation methods and query-scope conditions. complete hyperparameter details are provided in the appendix A.

5 RESULTS AND DISCUSSION

5.1 DOES RCD IMPROVE DENSE RETRIEVAL UNDER LIMITED SUPERVISION?

Table 2: Relative improvement in nDCG@8 over E5 Base across supervision budgets.

Method	10%	30%	50%	Avg. Gain
IMRNN Teacher	+4.3%	+10.6%	+17.3%	+10.8%
Supervised E5	+0.8%	+0.8%	+7.1%	+2.9%
Margin-MSE	-13.1%	-2.1%	+1.2%	-4.6%
EmbedDistill	-36.1%	-37.2%	-22.8%	-32.0%
RCD	**+6.9%**	**+13.3%**	**+21.7%**	**+14.0%**

Table 2 reveals three mechanisms behind the gains of RCD. With scarce labels, fine-tuning the full encoder yields almost no improvement, whereas an adapter learns far more from the same labels because it only models a correction to a fixed representation. RCD exploits this asymmetry, using the adapter to extract specialization from scarce labels and unlabeled queries to transfer it into the encoder. RCD's advantage over its teacher also persists as the teacher strengthens and the unlabeled pool shrinks, suggesting that a stronger teacher makes each unlabeled query more informative. Finally, under a weak teacher, Margin-MSE and EmbedDistill reproduce the teacher's outputs, inherit its errors, and fall below the base retriever. RCD instead aligns only the direction of the teacher's correction and lets ranking distillation set its extent, which explains its improvement even at the lowest supervision budget.

5.2 CAN RCD IMPROVE BEYOND ITS ADAPTER TEACHER?

Table 3: RCD performance at 50% supervision with an IMRNN teacher. We report RCD nDCG@8 and its relative change compared with the E5 base retriever, same-budget supervised E5, the frozen IMRNN teacher, Margin-MSE, and EmbedDistill.

Dataset	RCD	vs E5	vs Sup. E5	vs Teacher	vs Margin-MSE	vs EmbedDistill
ContractNLI	.192	+74.3%	-15.6%	+18.3%	+44.1%	+367.6%
PrivacyQA	.367	+46.6%	+24.7%	-0.7%	+13.5%	-7.2%
ComLQ	.829	+33.6%	+25.3%	+5.0%	+37.2%	+47.2%
SciFact	.770	+6.1%	+7.2%	+3.5%	+10.2%	+39.5%
NFCorpus	.365	-0.5%	+13.9%	-0.6%	+7.6%	+651.2%
Macro Avg.	**.504**	**+21.7%**	**+13.6%**	**+3.7%**	**+20.2%**	**+57.6%**

Table 3 shows that the student can exceed its own teacher. At 50% supervision, RCD outperforms the IMRNN teacher on ContractNLI, ComLQ, and SciFact and matches it on PrivacyQA and NFCorpus. This is possible because the teacher is trained only on the labeled queries Q_L, while the student learns the teacher's corrections on all queries in $Q_L \cup Q_U$ and therefore sees the specialization applied far more widely. The two ties fit this explanation. On NFCorpus, the teacher barely improves on the base retriever, so there is little to transfer. On PrivacyQA, the teacher improves substantially, but only 78 unlabeled queries are available to spread that improvement (see Table 1). The student thus surpasses its teacher when the teacher has learned a useful correction and enough unlabeled queries exist to propagate it, a condition we test directly in Section 5.5. The ComLQ example in Appendix C (Table 14) shows that RCD ranks relevant passages first, unlike competing methods that favor topically similar but irrelevant passages.

5.3 DOES RCD CAPTURE COMPOSITIONAL RELEVANCE?

Table 4: ComLQ performance across different logical query structures and hop counts at 50% supervision. We report nDCG@8 with approximate standard deviations across three seeds.

Subset	n	E5 Base	IMRNN	Margin-MSE	EmbedDistill	RCD
Overall	290	.619	$.789 \pm .004$	$.604 \pm .002$	$.563 \pm .010$	$\mathbf{.829} \pm .006$
Logical structure						
Negation	104	.510	$.784 \pm .010$	$.506 \pm .028$	$.450 \pm .029$	$\mathbf{.808} \pm .013$
Conjunction	87	.628	$.742 \pm .019$	$.617 \pm .009$	$.610 \pm .020$	$\mathbf{.800} \pm .010$
Union	41	.630	$.821 \pm .018$	$.612 \pm .040$	$.571 \pm .029$	$\mathbf{.841} \pm .007$
Projection	179	.615	$.802 \pm .004$	$.602 \pm .048$	$.566 \pm .048$	$\mathbf{.838} \pm .010$
Number of hops						
1-hop	15	.951	$.951 \pm .014$	$.877 \pm .065$	$.755 \pm .098$	$\mathbf{.975} \pm .000$
2-hop	211	.595	$.795 \pm .005$	$.580 \pm .016$	$.558 \pm .006$	$\mathbf{.821} \pm .007$
3-hop	64	.621	$.731 \pm .032$	$.618 \pm .044$	$.533 \pm .030$	$\mathbf{.818} \pm .007$

Dense retrievers struggle with queries that combine conditions through negation, conjunction, or multiple hops, because a single embedding tends to reflect the topics a query mentions rather than the logical relation among them. Such queries test whether a correction learned from limited labels captures relevance conditions that semantic similarity misses, or merely sharpens topical matching. ComLQ supports this test because it annotates each query by logical structure and hop count (Table 4). The base retriever performs worst on negation, where a query and a passage can share every topic yet disagree on the condition that determines relevance. RCD produces its largest improvement on this subset and achieves the best score on every logical structure, which suggests that the transferred correction encodes relevance conditions rather than topical similarity alone. The hop-count results reveal a second effect. The teacher's performance declines on three-hop queries, whereas RCD maintains its performance, suggesting that learning the correction over a larger set of queries yields a retriever that generalizes better to the most complex compositions. Margin-MSE, by contrast, falls below the base retriever on every subset. Matching the teacher's score margins therefore fails to transfer this compositional behavior, whereas aligning the teacher's correction succeeds.

5.4 WHY DOES RCD OUTPERFORM CONVENTIONAL DISTILLATION METHODS?

EmbedDistill performs below both RCD and the base retriever across supervision budgets. This pattern persists across different adapter architectures and base retrievers, with similar results observed using Search-Adaptor in place of IMRNNs and Qwen in place of E5. EmbedDistill uses Euclidean distance to match the complete adapted representation, requiring the student to reproduce both its direction and magnitude. RCD instead isolates the adapter-induced change and transfers its direction. We further isolate the effects of full representation distillation through an ablation in Section 5.6 (Table 6). Margin-MSE performs better than EmbedDistill, although the gap between Margin-MSE and RCD varies across datasets. At 50% supervision, RCD improves over Margin-MSE by 44.1% on ContractNLI, 13.5% on PrivacyQA, and 37.2% on ComLQ, compared with 10.2% on SciFact and 7.6% on NFCorpus (Table 3). Margin-MSE distills pairwise score margins between a preferred document and a negative document (Hofstätter et al., 2020). In our qrel-free setting, we use the teacher's highest-scoring candidate as the preferred document and compare it against the remaining candidates. This objective supervises individual document pairs without directly modeling relationships across the candidate set. RCD combines representation-level adaptation signals with listwise ranking distillation, which may be particularly useful when retrieval requires evidence from multiple documents. The ComLQ results are consistent with this explanation: the gap between RCD and Margin-MSE increases from approximately 11% on 1-hop queries to 42% and 32% on 2-hop and 3-hop queries, respectively (Table 4).

5.5 DOES UNLABELED TARGET DATA HELP PROPAGATE THE LEARNED SPECIALIZATION?

Table 5: Effect of unlabeled target queries on RCD at 50% supervision. All Queries uses both labeled and unlabeled target queries during student training, while Labeled Only restricts student training to the labeled subset.

Dataset	Labeled Only	All Queries	Δ	Rel. Gain
ContractNLI	0.175 ± .002	**0.192 ± .009**	+0.016	+9.3%
PrivacyQA	0.346 ± .005	**0.367 ± .002**	+0.020	+5.8%
ComLQ	0.810 ± .008	**0.829 ± .006**	+0.019	+2.3%
SciFact	0.742 ± .009	**0.770 ± .002**	+0.028	+3.4%
NFCorpus	0.353 ± .005	**0.365 ± .003**	+0.012	+3.3%

Table 5 compares RCD-AllQueries, which trains the student on $Q_L \cup Q_U$, with RCD-LabeledOnly, which restricts student training to Q_L, at 50% supervision. Both settings use the same labeled subset, frozen IMRNN teacher, and optimization budget. Across the datasets currently evaluated, using the additional unlabeled queries improves nDCG@8 in every case. Since the two settings use the same relevance labels to train the adapter teacher, the improvement comes from exposing the student to additional target-query texts rather than additional labels. For queries in Q_U, the frozen adapter provides representation-correction and ranking signals that extend the specialization learned from Q_L to queries that were not used to train the adapter. This supports the semi-supervised premise of RCD: unlabeled target queries provide additional opportunities to transfer the adapter's learned specialization into the standalone retriever.

5.6 WHAT DOES EACH RCD SIGNAL CONTRIBUTE?

Table 6 examines the contribution of each distillation signal and the choice of representation target. Combining correction alignment with ranking distillation achieves the highest macro-average nDCG@8, although the individual objectives perform better on some datasets. As discussed in Section 5.4, EmbedDistill differs from RCD in both its representation target and its use of Euclidean distance. To isolate the effect of the representation target, RCD (Final Representation + Ranking) uses cosine similarity to align the final representations while retaining the same ranking objective. Its lower macro-average nDCG@8 (.411 versus .505) provides further evidence that explicitly aligning adapter-induced corrections helps transfer the learned specialization. The difference is particularly significant on NFCorpus, both EmbedDistill (Table 3) and the RCD final-representation ablation (Table 6) exhibit severe performance degradation, despite using different alignment objectives. Their consistently poor performance suggests that directly matching complete teacher representations may

Table 6: Ablation of the RCD supervision signals and representation target at 50% supervision with an IMRNN teacher. Ranking Only uses ranking distillation, Correction Only uses representation correction, RCD (Final Rep. + ranking) aligns final teacher representations together with ranking distillation, and RCD combines correction alignment with ranking distillation.

Dataset	Ranking Only	Correction Only	RCD (Final Rep. + ranking)	RCD
ContractNLI	.135 ± .014	.166 ± .008	.160 ± .018	**.192 ± .009**
PrivacyQA	.243 ± .014	.309 ± .013	.345 ± .009	**.367 ± .002**
ComLQ	**.844 ± .004**	.734 ± .004	.803 ± .005	.829 ± .006
SciFact	.732 ± .005	.759 ± .010	.730 ± .007	**.770 ± .002**
NFCorpus	.363 ± .001	**.380 ± .002**	.017 ± .001	.365 ± .003
Macro Avg.	.463	.470	.411	**.505**

distort the relative query-document similarities learned by the student, even when combined with ranking distillation.

5.7 Does RCD Generalize Across Adapter Architectures?

Table 7: Generalization to Search-Adaptor teachers across supervision budgets. We report the relative change in nDCG@8 over E5 Base.

Method	10%	30%	50%
Search-Adaptor	-0.6%	+7.3%	**+9.6%**
Margin-MSE	-0.5%	+6.2%	+9.2%
EmbedDistill	-30.2%	-11.3%	-8.0%
RCD	**+0.9%**	**+8.3%**	+5.6%

The preceding experiments use IMRNNs as the adapter teacher. We next replace IMRNNs with Search-Adaptor to examine whether RCD depends on a particular adapter architecture. Table 7 shows that RCD continues to improve over the pretrained E5 retriever across all three supervision budgets, with its largest average gain of 8.3% at 30% supervision. At the lower two budgets, RCD also improves upon the corresponding Search-Adaptor teacher. The behavior at 50% supervision is less uniform, where the Search-Adaptor teacher and Margin-MSE perform better than RCD. Nevertheless, RCD remains above the pretrained retriever at every supervision level and consistently outperforms direct embedding distillation. Together with the IMRNN experiments, these results indicate that RCD is not specific to a single adapter architecture, although the magnitude of the benefit depends on the teacher and supervision level.

6 Conclusion

This work changes the role of retriever adapters. Prior work treats a trained adapter as the endpoint of adaptation; we, instead, show that it can teach its own base retriever, turning specialization learned from scarce labels into supervision for unlabeled target queries. RCD realizes this transfer and yields a standalone retriever that improves on its base model and can surpass the adapter that taught it. The study also shifts what retrieval distillation should transfer. Existing methods distill a teacher's scores or final representations. When teacher and student share a base retriever, however, the teacher's specialization resides in how it changes the base representation, and distilling the full representation obscures that change. Aligning the direction of the change, together with the teacher's ranking behavior, transfers the specialization, whereas matching complete representations can degrade retrieval. Finally, the results reframe label-efficient adaptation. Relevance labels remain necessary to learn what relevance means in a new domain, but unlabeled queries determine how far that knowledge spreads. A student trained only on labeled queries matches its teacher, whereas one trained on additional unlabeled queries exceeds it. For practitioners, this makes unlabeled query logs, which accumulate at no labeling cost, a direct source of retrieval improvement. These findings suggest that the lightweight adapters already used to specialize retrievers can also be used to improve the retrievers themselves.

AI ASSISTANCE

We used large language models to improve the clarity and readability of the manuscript. The authors take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 describes RCD and its training objective, while Section 4 specifies the datasets, supervision budgets, baselines, and evaluation protocol. Appendix A provides the implementation details and training hyperparameters for the student and adapter teachers. Additional retrieval results and qualitative examples are reported in Appendices B and C, respectively. The implementation and scripts for reproducing our experiments are available in an anonymized repository at `https://anonymous.4open.science/r/RCD-F375/`.

REFERENCES

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2387–2392, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531863. URL https://doi.org/10.1145/3477495.3531863.

Ramraj Chandradevan, Kaustubh Dhole, and Eugene Agichtein. DUQGen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7437–7451, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.413. URL https://aclanthology.org/2024.naacl-long.413/.

Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pp. 95–110, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-030-99735-9. doi: 10.1007/978-3-030-99736-6_7. URL https://doi.org/10.1007/978-3-030-99736-6_7.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*, 2022.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 113–122, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462891. URL https://doi.org/10.1145/3404835.3462891.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Chao-Wei Huang and Yun-Nung Chen. PairDistill: Pairwise relevance distillation for dense retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pp. 18225–18237, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1013. URL https://aclanthology.org/2024.emnlp-main.1013/.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL https://arxiv.org/abs/2112.09118.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. Embeddistill: A geometric knowledge distillation for information retrieval, 2023. URL https://arxiv.org/abs/2301.12005.

Dayoon Ko, Jinyoung Kim, Sohyeon Kim, Jinhyuk Kim, Jaehoon Lee, Seonghak Song, Minyoung Lee, and Gunhee Kim. When should dense retrievers be updated in evolving corpora? detecting out-of-distribution corpora using GradNormIR. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25977–25996, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1334. URL https://aclanthology.org/2025.findings-acl.1334/.

Dohyeon Lee, Jongyoon Kim, Jihyuk Kim, Seung-won Hwang, and Joonsuk Park. tRAG: Term-level retrieval-augmented generation for domain-adaptive retrieval. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6566–6578, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.334. URL https://aclanthology.org/2025.naacl-long.334/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, and Nan Duan. Prod: Progressive distillation for dense retrieval. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 3299–3308, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583421. URL https://doi.org/10.1145/3543507.3583421.

Tao Liu, Chenshu Chen, Xi Yang, and Wenming Tan. Rethinking knowledge distillation with raw features for semantic segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1144–1153, 2024. doi: 10.1109/WACV57701.2024.00119.

Seiji Maekawa, Moin Aminnaseri, Pouya Pezeshkpour, and Estevam Hruschka. Align then train: Efficient retrieval adapter learning, 2026. URL https://arxiv.org/abs/2604.03403.

Roxana Petcu, Samarth Bhargav, Maarten de Rijke, and Evangelos Kanoulas. A comprehensive taxonomy of negation for NLP and neural retrievers. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15511–15533, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.839. URL https://aclanthology.org/2025.findings-emnlp.839/.

Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain, 2024. URL https://arxiv.org/abs/2408.10343.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.224. URL https://aclanthology.org/2021.emnlp-main.224/.

Joseph John Rocchio Jr. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, 1971.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11265–11279, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.693. URL https://aclanthology.org/2023.emnlp-main.693/.

Suvansh Sanjeev and Anton Troynikov. Embedding adapters, May 2024. URL https://www.trychroma.com/research/embedding-adapters.

Yash Saxena, Ankur Padia, Kalpa Gunaratna, and Manas Gaur. IMRNNs: An efficient method for interpretable dense retrieval via embedding modulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6324–6337, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.333. URL https://aclanthology.org/2026.findings-eacl.333/.

Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10186–10204, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.511. URL https://aclanthology.org/2025.naacl-long.511/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf.

Junjie Wang, Yicheng Chen, Wangshu Zhang, Sen Hu, Teng Xu, and Jing Zheng. AdapterDistillation: Non-destructive task composition with knowledge distillation. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 194–201, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.20. URL https://aclanthology.org/2023.emnlp-industry.20/.

Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In Marie-Francine Moens,

Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.59. URL https://aclanthology.org/2021.findings-emnlp.59/.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2345–2360, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.168. URL https://aclanthology.org/2022.naacl-main.168/.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022b.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4008–4020, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.316. URL https://aclanthology.org/2022.findings-acl.316/.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Ganlin Xu, Zhitao Yin, Linghao Zhang, Jiaqing Liang, Weijia Lu, Xiaodong Zhang, Zhifei Yang, Sihang Jiang, and Deqing Yang. Comlq: Benchmarking complex logical queries in information retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34115–34123, 2026.

Eugene Yang, Suraj Nair, Dawn Lawrie, James Mayfield, and Douglas W Oard. Parameter-efficient zero-shot transfer for cross-language dense retrieval with adapters. *arXiv preprint arXiv:2212.10448*, 2022.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. Search-adaptor: Embedding customization for information retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12230–12247, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.661. URL https://aclanthology.org/2024.acl-long.661/.

Jinsung Yoon, Rajarishi Sinha, Sercan O Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10318–10336, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.576. URL https://aclanthology.org/2024.emnlp-main.576/.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1462–1479, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.95. URL https://aclanthology.org/2022.emnlp-main.95/.

Sepanta Zeighami, Zac Wellmer, and Aditya Parameswaran. Nudge: Lightweight non-parametric fine-tuning of embeddings for retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 21528–21560, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

A Implementation Details

A.1 Training Hyperparameters

Unless otherwise stated, we use the same student distillation configuration across backbone models and teacher architectures. In particular, the hyperparameters in Table 8 are shared across experiments with E5 and Qwen3-Embedding, and with IMRNNs and Search-Adaptor teachers. Only the underlying backbone and teacher are changed in the corresponding generalization experiments.

Table 8: Student distillation configurations used throughout the experiments. The same settings are used across backbone models and teacher architectures, unless otherwise stated.

Setting	**Margin-MSE**	**EmbedDistill**	**RCD**
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Weight decay	0.01	0.01	0.01
Maximum epochs	5	5	5
Warmup ratio	0.05	0.05	0.05
Gradient clipping	1.0	1.0	1.0
Candidate pool size	64	64	64
Training candidates / query	32	32	32
Query batch size	1	1	1
Evaluation batch size	32	32	32
Early-stopping patience	3	3	3
Mixed precision	Yes	Yes	Yes
Student training uses relevance labels	No	No	No

A.2 Teacher-Specific Implementation Details

IMRNNs. We use the public IMRNNs implementation (Saxena et al., 2026) for the primary adapter teacher.[1] For each supervision budget, the IMRNN teacher is trained only on the corresponding labeled subset and is frozen before student distillation. We train the teacher for up to 10 epochs using a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, batch size 32, and 32 negatives per query. The underlying dense retriever remains frozen during adapter training. Other architectural and implementation details follow the released IMRNNs implementation.

Search-Adaptor. For the teacher-architecture generalization experiments, we use Search-Adaptor (Yoon et al., 2024a). We follow its ranking, recovery, and prediction objectives and use Adam with a learning rate of 10^{-3} and batch size 128. Training is performed for at most 2,000 iterations with an early-stopping patience of 125 iterations. Negative document pairs are subsampled at a ratio of 10. The recovery and prediction weights are selected on the validation set using $\alpha \in \{0, 0.1, 1.0\}$ and $\beta \in \{0, 0.01, 0.1\}$.

The Search-Adaptor paper specifies MLP-based adaptation and prediction modules but does not specify all architectural choices needed for reimplementation. We therefore use a two-layer ReLU MLP with a residual connection, shared query and document adapter parameters, and a hidden width equal to the embedding dimension of the underlying retriever. The query predictor is used only during teacher training. We do not use dropout or layer normalization. To keep the comparison with IMRNNs controlled, Search-Adaptor uses the same benchmark train, validation, and test splits and the same labeled subsets at each supervision budget.

A.3 Computational Cost

We report wall-clock student training time under a matched setting on NFCorpus with a 50% relevance-supervision budget. All methods use the same IMRNN teacher, frozen candidate pool, and a single NVIDIA L40S GPU. Training time is measured from student optimizer initialization to the training-completion checkpoint and therefore includes optimization, validation, and checkpointing between epochs, while excluding teacher training and one-time candidate construction.

[1] `https://pypi.org/project/imrnns/`

Table 9: Wall-clock student training time on NFCorpus with a 50% relevance-supervision budget. All methods use the same IMRNN teacher and are run on a single NVIDIA L40S GPU.

Method	**Epochs**	**Training Time**
Margin-MSE	4	1 h 30 min
EmbedDistill	5	1 h 53 min
RCD	5	1 h 53 min

RCD has a training cost comparable to EmbedDistill in this setting. After distillation, RCD does not retain the teacher or introduce additional inference-time modules; retrieval is performed directly by the standalone student encoder.

B Additional Retrieval Results

This section reports results omitted from the main text for space. We first evaluate retrieval at different cutoffs, then provide the full Search-Adaptor results across supervision budgets, and finally report the Qwen3-Embedding experiments.

B.1 Results Across Retrieval Cutoffs

Table 10: Test-set nDCG across retrieval cutoffs at 50% supervision with an IMRNN teacher. Results are the mean ± standard deviation across available seeds. IMRNN (Teacher) and E5 (Base: Student)

Dataset	Method	nDCG@2	nDCG@4	nDCG@8	nDCG@16	nDCG@32	nDCG@64
ContractNLI	IMRNN Teacher	.090 ± .008	.111 ± .012	.162 ± .006	.159 ± .013	.183 ± .012	.204 ± .008
	Margin-MSE	.086 ± .009	.092 ± .016	.133 ± .012	.119 ± .013	.127 ± .009	.136 ± .011
	EmbedDistill	.000 ± .000	.000 ± .000	.041 ± .002	.031 ± .002	.016 ± .004	.027 ± .002
	Ranking Only	.108 ± .029	.131 ± .027	.135 ± .014	.158 ± .030	.168 ± .032	.178 ± .034
	Correction Only	.111 ± .011	.143 ± .007	.166 ± .008	.176 ± .009	.192 ± .011	.205 ± .011
	RCD	.130 ± .022	.144 ± .027	.192 ± .009	.174 ± .023	.185 ± .024	.198 ± .022
PrivacyQA	IMRNN Teacher	.413 ± .021	.382 ± .007	.370 ± .015	.451 ± .017	.498 ± .017	.529 ± .005
	Margin-MSE	.314 ± .035	.320 ± .022	.323 ± .027	.376 ± .041	.406 ± .038	.444 ± .037
	EmbedDistill	.361 ± .010	.372 ± .005	.395 ± .005	.399 ± .017	.422 ± .017	.456 ± .020
	Ranking Only	.204 ± .049	.233 ± .033	.243 ± .014	.264 ± .019	.301 ± .011	.335 ± .012
	Correction Only	.375 ± .010	.340 ± .013	.309 ± .013	.427 ± .008	.487 ± .015	.531 ± .008
	RCD	.377 ± .035	.366 ± .016	.367 ± .002	.432 ± .021	.483 ± .017	.532 ± .018
ComLQ	IMRNN Teacher	.701 ± .001	.754 ± .001	.790 ± .004	.815 ± .003	.822 ± .002	.823 ± .002
	Margin-MSE	.592 ± .005	.599 ± .002	.604 ± .002	.644 ± .003	.678 ± .002	.701 ± .003
	EmbedDistill	.475 ± .005	.533 ± .003	.563 ± .010	.601 ± .001	.620 ± .001	.631 ± .001
	Ranking Only	.764 ± .012	.808 ± .009	.844 ± .004	.856 ± .005	.860 ± .005	.862 ± .006
	Correction Only	.573 ± .010	.661 ± .008	.734 ± .004	.738 ± .008	.741 ± .008	.742 ± .008
	RCD	.732 ± .003	.789 ± .001	.829 ± .006	.843 ± .001	.847 ± .001	.847 ± .001
SciFact	IMRNN Teacher	.690 ± .003	.724 ± .002	.744 ± .001	.757 ± .001	.763 ± .001	.770 ± .001
	Margin-MSE	.709 ± .010	.743 ± .007	.699 ± .005	.771 ± .004	.782 ± .004	.786 ± .001
	EmbedDistill	.494 ± .017	.530 ± .013	.552 ± .009	.566 ± .007	.578 ± .006	.589 ± .002
	Ranking Only	.693 ± .011	.728 ± .007	.732 ± .005	.764 ± .007	.770 ± .007	.777 ± .005
	Correction Only	.664 ± .009	.693 ± .010	.759 ± .010	.733 ± .007	.742 ± .006	.747 ± .007
	RCD	.661 ± .005	.694 ± .010	.770 ± .002	.736 ± .006	.743 ± .006	.750 ± .006
NFCorpus	IMRNN Teacher	.320 ± .010	.327 ± .008	.367 ± .003	.351 ± .004	.368 ± .007	.393 ± .002
	Margin-MSE	.299 ± .001	.312 ± .002	.339 ± .002	.315± .009	.317 ± .007	.314 ± .001
	EmbedDistill	.014 ± .005	.010 ± .006	.049 ± .005	.031 ± .003	.027 ± .003	.033 ± .001
	Ranking Only	.404 ± .001	.390 ± .003	.363 ± .001	.338 ± .000	.321 ± .001	.319 ± .001
	Correction Only	.432 ± .001	.410 ± .002	.380 ± .002	.350 ± .002	.334 ± .001	.330 ± .001
	RCD	.416 ± .003	.389 ± .005	.365 ± .003	.334 ± .001	.318 ± .001	.319 ± .000

Table 10 reports the same 50% supervision setting used in the main analysis at additional retrieval cutoffs. The relative ordering varies with k on some datasets. RCD is strongest at nDCG@8 on ContractNLI, ComLQ, and SciFact, while other objectives can perform better at smaller or larger cutoffs. NFCorpus is less stable across cutoffs: correction-only variants perform better at small k, while the differences narrow as more documents are retrieved.

B.2 Results Across Supervision Budgets

Table 11 gives the dataset-level results underlying the Search-Adaptor averages reported in Section 5.7. The teacher, Margin-MSE, EmbedDistill, and RCD use the same labeled subset at each supervision budget. As with the IMRNN experiments, student distillation does not use relevance labels.

Table 11: Test-set nDCG@8 using Search-Adaptor as the teacher across different supervision budgets.

Supervision	Dataset	Search-Adaptor	Margin-MSE	EmbedDistill	RCD
10%	ContractNLI	.121	.140	.076	.131
	PrivacyQA	.254	.277	.240	.254
	ComLQ	.641	.604	.603	.653
	SciFact	.725	.723	.431	.728
	NFCorpus	.320	.319	.097	.326
30%	ContractNLI	.134	.153	.137	.134
	PrivacyQA	.388	.375	.415	.392
	ComLQ	.639	.627	.658	.649
	SciFact	.729	.722	.495	.733
	NFCorpus	.334	.325	.134	.337
50%	ContractNLI	.127	.136	.126	.132
	PrivacyQA	.424	.426	.480	.400
	ComLQ	.636	.638	.648	.613
	SciFact	.736	.731	.501	.703
	NFCorpus	.349	.333	.152	.341
100%	ContractNLI	.130	.124	.041	.140
	PrivacyQA	.440	.406	.397	.473
	ComLQ	.634	.596	.560	.655
	SciFact	.739	.725	.434	.791
	NFCorpus	.377	.354	.098	.377

The gains vary substantially by dataset. RCD improves over the E5 base retriever at each supervision budget on average, but it does not consistently outperform the Search-Adaptor teacher. At 50% supervision, for example, RCD improves most clearly on ComLQ, while the teacher performs better on PrivacyQA and SciFact. EmbedDistill remains below the other distillation methods on most settings.

B.3 Generalization to Qwen3-Embedding

We also replace E5-large-v2 with Qwen3-Embedding while keeping the IMRNN teacher construction and student-training protocol unchanged. Table 12 reports the dataset-level results using a 50% relevance-supervision budget.

Table 12: Generalization to Qwen3-Embedding with a 50% relevance-supervision budget. We report mean nDCG@8 over three optimization seeds, with standard deviation shown after $\pm$.

Method	ContractNLI	PrivacyQA	ComLQ	SciFact	NFCorpus	Macro Avg.
Qwen3-Embedding Base	.131	.176	.715	.698	**.366**	.417
Supervised Qwen, 50%	**.320 ± .010**	.250 ± .006	.592 ± .003	.699 ± .016	.355 ± .007	.443
IMRNN-Qwen Teacher, 50%	.191 ± .003	.303 ± .001	**.789 ± .001**	.727 ± .001	.365 ± .001	.475
Margin-MSE	.160 ± .007	.274 ± .006	.750 ± .005	.706 ± .002	.184 ± .006	.415
EmbedDistill	.069 ± .019	.242 ± .009	.675 ± .003	.486 ± .018	.152 ± .007	.325
RCD	.169 ± .007	**.359 ± .014**	.761 ± .003	**.784 ± .002**	.346 ± .002	**.484**

RCD improves the Qwen3-Embedding base on four of the five datasets and raises macro-average nDCG@8 from .417 to .484. The exception is NFCorpus, where the base retriever scores .366 and RCD scores .346. The IMRNN teacher remains stronger than RCD on ComLQ and NFCorpus, while RCD exceeds it on PrivacyQA and SciFact.

Table 13 breaks down the Qwen3-Embedding results on ComLQ. The relative behavior differs from the E5 experiments: the IMRNN teacher remains stronger than RCD on several subsets, although RCD stays above the Qwen base on most logical structures and multi-hop subsets.

Table 13: ComLQ breakdown with Qwen3-Embedding with a 50% relevance-supervision budget. We report nDCG@8 by logical operator and hop count. Results are mean over three optimization seeds, with standard deviation shown after $\pm$. n denotes the number of test queries.

Category	Qwen Base	IMRNN Teacher	Margin-MSE	EmbedDistill	RCD
Overall ($n = 290$)	.715	**.789 ± .001**	.750 ± .005	.675 ± .003	.761 ± .003
Negation ($n = 104$)	.656	**.769 ± .002**	.736 ± .010	.669 ± .004	.740 ± .005
Conjunction ($n = 87$)	.712	**.775 ± .002**	.739 ± .008	.646 ± .010	.742 ± .002
Union ($n = 41$)	.729	**.791 ± .004**	.719 ± .015	.686 ± .006	.775 ± .008
Projection ($n = 179$)	.720	**.787 ± .001**	.745 ± .007	.678 ± .003	.754 ± .001
1-hop ($n = 15$)	.902	**.951 ± .001**	.942 ± .020	.853 ± .015	.896 ± .004
2-hop ($n = 211$)	.698	**.775 ± .001**	.729 ± .004	.663 ± .003	.748 ± .003
3-hop ($n = 64$)	.727	**.797 ± .003**	.772 ± .006	.672 ± .006	.770 ± .005

C Qualitative Examples

Table 14 shows three test queries where the retrieved passages make the differences between the methods easy to inspect. For each method, we show its top-ranked passage and whether that passage is labeled relevant. We also report how many of the query's ground-truth relevant passages appear in the top eight. These examples are selected for analysis and are not intended to represent average performance.

The examples show two types of improvement. On ComLQ and ContractNLI, the competing methods place lexically or topically related passages at rank 1, while RCD retrieves a passage that is labeled relevant to the query. On PrivacyQA, both the IMRNN teacher and RCD retrieve a relevant passage at rank 1, but RCD retrieves three of the four relevant passages within the top eight, compared with one for the teacher.

D Limitations and Future Work

RCD currently applies to settings where the base model, teacher, and student operate in a shared embedding space, since the representation-correction signal is defined as the change relative to the base representation. RCD also depends on the quality of the teacher signal. Although the student can outperform its teacher, the specialization learned by the teacher determines the supervision available for transfer and can therefore limit the gains achievable by the student. While we study retriever adapters as teachers, RCD itself is not limited to adapter-based specialization. A specialized version of a shared base model can potentially serve as the teacher as long as its representation changes relative to the base model can be meaningfully measured. This could, for example, include a domain-specialized retriever obtained through direct fine-tuning rather than through an adapter. RCD also has the potential to be used in other dense-retrieval adaptation settings, such as cross-lingual and continual embedding adaptation. Exploring these settings, as well as extending RCD to models operating in different embedding spaces, provides promising directions for future work.

Table 14: Selected retrieval examples at 50% supervision. For each query, we show the top-ranked passage from each method, whether it is labeled relevant, the number of ground-truth relevant passages retrieved in the top eight, and nDCG@8. Ellipses indicate omitted passage text.

Method	Top-1 retrieved passage	Top-1 relevant?	Relevant in top 8	nDCG@8
ComLQ				
Query: Which individuals contributed to the continuation or expansion of ecclesiastical history? *Ground truth: 2 relevant passages.* Representative evidence: Jerome's work encouraged later chroniclers, including **Prosper, Cassiodorus, and Victor of Tunnuna**, to continue his annals.				
E5	"Eusebius of Caesarea is renowned for his meticulous cataloging of influential figures … Tertullian, an early Christian apologist and theologian …"	✗	1/2 (3)	0.307
IMRNN teacher	"… theological and cultural transitions which took place between the Old and New Testaments … apocryphal, deuterocanonical, or intertestamental books …"	✗	2/2 (4, 7)	0.468
Margin-MSE	Same Eusebius–Tertullian passage as E5.	✗	1/2 (5)	0.237
EmbedDistill	Same Eusebius–Tertullian passage as E5.	✗	1/2 (3)	0.307
RCD	"… Jerome produced a valuable work … later chroniclers as **Prosper, Cassiodorus, and Victor of Tunnuna to continue his annals**."	✓	**2/2 (1, 2)**	**1.000**
ContractNLI				
Query: Does JB Machine LLC's NDA permit the receiving party to copy confidential information under certain circumstances? *Ground truth: 2 relevant passages.* Representative evidence: The recipient must **not copy or modify information without the owner's prior written assent**.				
E5	"NON-DISCLOSURE AGREEMENT … This Non-Disclosure Agreement … was freely stipulated and entered into … between: JB Machine LLC …"	✗	0/2	0.000
IMRNN teacher	"A person's memory is unaided if the person has not intentionally memorized the Confidential Information …" (From a different NDA.)	✗	0/2	0.000
Margin-MSE	Same JB Machine NDA opening as E5.	✗	0/2	0.000
EmbedDistill	"gnatory]" (A truncated fragment from a different NDA.)	✗	0/2	0.000
RCD	"**Not copy or, in any way, modify any information without the prior written ascent of the Owner** …"	✓	**1/2 (1)**	**0.613**
PrivacyQA				
Query: Does 23andMe anonymize my data? *Ground truth: 4 relevant passages.* Representative evidence: Registration information is removed from sensitive information and the data is assigned **a randomly generated ID** to prevent reasonable identification.				
E5	"23andMe does not endorse or sponsor any API applications, and does not affirm the accuracy or validity of any interpretations made by third party API applications …"	✗	1/4 (5)	0.151
IMRNN teacher	"De-identification/Pseudonymization. **Registration Information is stripped from Sensitive Information** … This data is then assigned **a randomly generated ID** …"	✓	1/4 (1)	0.390
Margin-MSE	"23andMe may share summary statistics, which do not identify any particular individual … with our qualified research collaborators."	✗	0/4	0.000
EmbedDistill	"Under certain circumstances your Personal Information may be subject to processing pursuant to laws, regulations, judicial or other government subpoenas …"	✗	0/4	0.000
RCD	"De-identification/Pseudonymization. **Registration Information is stripped from Sensitive Information** … This data is then assigned **a randomly generated ID** …"	✓	**3/4 (1, 6, 8)**	**0.653**

Reading the table: For each query, the ground-truth count gives the total number of labeled relevant passages. Relevant-in-top-8 is reported as retrieved/total, with the retrieved passages' ranks in parentheses. For example, 3/4 (1, 6, 8) means that three of four relevant passages appear at ranks 1, 6, and 8. Checkmarks indicate that the top-ranked passage is labeled relevant; crosses indicate that it is not.