UnIte: Uncertainty-based Iterative Document Sampling for Domain Adaptation in Information Retrieval

Jongyoon Kim Minseong Hwang Seung-won Hwang*
Interdisciplinary Program in Artificial Intelligence, Seoul National University
{john.jongyoon.kim, hwmin0823, seungwonh}@snu.ac.kr

Abstract

Unsupervised domain adaptation generalizes neural retrievers to an unseen domain by generating pseudo queries on target domain documents. The quality and efficiency of this adaptation critically depend on which documents are selected for pseudo query generation. The existing document sampling method focuses on diversity but fails to capture model uncertainty. In contrast, we propose **Un**certainty-based **Ite**rative Document Sampling (UnIte) addressing these limitations by (1) filtering documents with high aleatoric uncertainty and (2) prioritizing those with high epistemic uncertainty, maximizing the learning utility of the current model. We conducted extensive experiments on a large corpus of BEIR with small and large models, showing significant gains of +2.45 and +3.49 nDCG@10 with a smaller training sample size, 4k on average. [1]

1 Introduction

Neural retrievers pre-trained on large datasets, such as MS-MARCO, achieve strong performance in the seen domain while being limited to unseen domains (Thakur et al., 2021). Unsupervised Domain Adaptation (UDA) via pseudo query generation has emerged as a promising solution, finetuning the retriever on target-domain documents paired with generated queries (Thakur et al., 2021; Ma et al., 2021). Yet for corpora exceeding 100k documents, the number of generator calls scales with corpus size and is often infeasible under typical budgets (Gospodinov et al., 2023). Therefore, sampling *which* documents to query becomes the central bottleneck.

Unlike prior works that randomly sample documents (Wang et al., 2022; Bonifacio et al., 2022; Dai et al., 2022), DUQGen samples with a diversity-based approach (Chandradevan et al., 2024), with an external embedding model[2]. While diversity improves coverage, not all regions contribute equally to adaptation. We observe two characteristic regions in DUQGen's sampling as illustrated on Figure 1: (i) **low-density (red-circled region)** regions where it contains atypical or outlier documents, and (ii) **high-confidence (orange-circled)** regions where the model's predictive uncertainty is low (already-learned areas), yielding a limited learning signal.

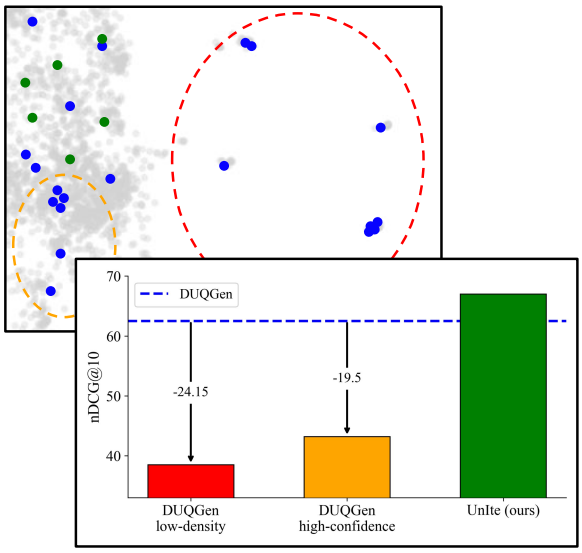

Figure 1: The scatter plot shows that DUQGen tends to select samples from low-density (red) or high-confidence (orange) regions in TREC-COVID with Contriever (cropped for visibility). These samples lead to performance degradation, indicating they are suboptimal for domain adaptation.

Training exclusively on each region drops performance by 24.15 (red bar) and 19.32 (orange bar) nDCG@10 relative to DUQGen (62.56), as shown in Figure 1, indicating a limited learning signal for adaptation, while sampled heavily (5% and

* Corresponding Authors

[1] The implementation is available at: https://github.com/ldilab/UnIte.

[2] Contriever is used for sampling.

Findings of the Association for Computational Linguistics: ACL 2026, pages 32256–32272

13%). Meanwhile, according to uncertainty taxonomy (Hüllermeier and Waegeman, 2021), low-density regions correspond to high Aleatoric Uncertainty (AU), that is inherent in ambiguous or noisy data, while high-confidence areas correspond to low Epistemic Uncertainty (EU). EU reflects the model's knowledge gaps, that is, documents with high EU are misaligned with the target domain and thus informative for adaptation. This suggests prioritizing *low-AU, high-EU* documents while maintaining diversity for coverage.

We propose **Un**certainty-based **Ite**rative Document Sampling **(UnIte)**, which estimates and exploits both forms of uncertainty for document sampling. (1) **AU (data).** We use a density proxy based on the lexical distance to the k-th nearest neighbor. If this neighbor is distant, the document is considered to lie in a low-density region. We filter out noisy documents in low-density regions to prevent selection failures. (2) **EU (model).** We measure how well the current retriever aligns a document with the target domain by comparing the document's embedding with a vocabulary-based domain distribution derived from the model. A poor document representation of important terms, such as having high frequency in the domain, is treated as high EU.

Another challenge is that EU shifts as the model adapts. As a result, good samples may become suboptimal after shifts. We therefore adopt an *iterative* sample–train loop that recomputes EU each round, and we stop when uncertainty plateaus.

We evaluate on five BEIR datasets with large corpora (>100k documents). UnIte improves over DUQGen by +2.45 on the small model, DPR, and +3.49 on large model, Qwen3-Embedding-4B, on average nDCG@10 while using fewer pseudo-queries. Ablations show that removing epistemic sampling drops nDCG@10 by 2.3, and an additional drop of 0.9 is caused by removing the aleatoric filter, underscoring the need for uncertainty-aware selection.

2 Related Work

This section synthesizes prior UDA works and studies the progress in document selection for UDA. We organize the discussion into three parts: (i) UDA with pseudo-query generation, (ii) document sampling strategies, and (iii) our distinction.

Unsupervised Domain Adaptation UDA for neural retrieval typically proceeds by generating pseudo-queries from target-domain documents and then fine-tuning a source-trained retriever, such as, one trained with MS-MARCO (Thakur et al., 2021; Wang et al., 2022). Early approaches used Doc2Query (Gospodinov et al., 2023), while recent ones employ LLM prompting with domain descriptions (Asai et al., 2023), contrastive (Bonifacio et al., 2022) or few-shot examples (Dai et al., 2022), or query expansion (Lee et al., 2025). Several approaches also filter low-quality queries using perplexity or consistency checks (Bonifacio et al., 2022; Dai et al., 2022). These methods improve pseudo-query quality, but generating queries for all documents is infeasible, leading most systems to randomly sample documents under a fixed budget.

Document Sampling Because pseudo queries are generated from sampled documents, gains from improving the quality of pseudo queries are fundamentally bounded by the sampling step. Consequently, recent UDA methods seek to improve the quality of the sampled documents themselves. Prominent directions in document sampling include (i) maximizing target-domain coverage via diversity-based selection (Chandradevan et al., 2024) and (ii) estimating document quality by training neural models to approximate perplexity or related signals (Chang et al., 2024; Lawrie et al., 2025). However, these approaches do not account for both the target-domain distribution and the current model knowledge on the target domain. For example, DUQGen (Chandradevan et al., 2024) relies on an external model (Contriever) that is irrelevant to the target retriever's learning state, and performs a single static selection that cannot adapt as the model trains. As a result, it often selects uninformative documents for adaptation, thereby wasting the selection budget.

Our Distinction Unlike prior work DUQGen (Chandradevan et al., 2024), which neglects the target-domain distribution and the model's current knowledge, we explicitly estimate uncertainty at both the data and model levels. These estimations drive our sampling policy, which prioritizes low-AU, high-EU documents with diversity to keep coverage, so that the selected set is tailored to the specific domain and retriever. This yields more informative updates per query-generation budget than random or purely diversity-based selection.

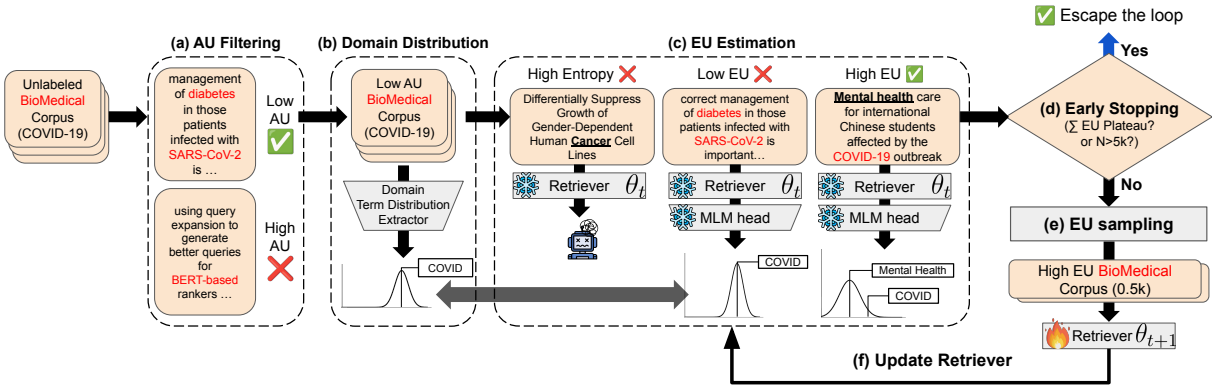

Figure 2: UnIte pipeline overview. An example from the Biomedical Domain (TREC-COVID) is illustrated. AU filtering is performed first on the corpus. Then, the sampling-training loop samples documents based on EU and iterates until the maximum budget is reached or meets early stopping criteria.

3 Methodology

As shown in Figure 2, this section explains how UnIte leverages uncertainties. The algorithmic details are provided in Appendix A.1.

3.1 AU Filtering

Figure 2-(a) illustrates a concrete example of this filtering process within the Biomedical (COVID-19) domain. The corpus contains some off-topic documents discussing "BERT-based rankers". Since this document is lexically distant from the dense region of biomedical terms, UnIte identifies its document d as having high AU and filters it out from the corpus $\mathcal{C}$, thereby preventing negative transfer.

Following Hacohen et al. (2022), we identify outliers in low-density regions as high-AU samples. Since AU is inherently a property of the data itself, independent of any particular model (Hüllermeier and Waegeman, 2021), its estimation must also be model-free. Using neural embeddings for this purpose would conflate data uncertainty with model uncertainty. A document may appear as an outlier simply because the model has not yet learned to represent it (high EU), rather than being inherently noisy (high AU). To ensure clean separation from EU, we adopt a lexical distance metric (Hu et al., 2019) based on the BM25 score, which relies solely on corpus statistics.

Formally, the distance to the k-th nearest neighbor n_k is defined as:

$$D_k(d) = \frac{1}{\epsilon + \text{BM25}(d, n_k)} \quad (1)$$

where ϵ, set to $1e-6$, prevents zero division. We normalize these distances using the modified z-score $z(\cdot)$ to ensure compatibility across domains. As illustrated in Figure 2-(a), off-topic documents, such as the one discussing "BERT-based rankers", exhibit high distances ($z(d) > z_{\text{thr}}$) due to the lack of shared terms with the corpus. These high AU documents are filtered out to yield a refined corpus $\mathcal{C}'$:

$$\mathcal{C}' = \{d \in \mathcal{C} \mid z(d) \leq z_{\text{thr}}\} \quad (2)$$

3.2 Iterative Sampling-Training Loop

A one-shot sampling strategy relies on a static assessment of document informativeness derived from the initial model θ_0, failing to capture the model's evolving understanding. Documents that were initially uncertain may become trivial over time, causing static selections to suffer from information redundancy (Settles, 2009; Ash et al., 2019). To address this, we employ an iterative sampling-training loop (Figure 2-(c–f)) that continuously re-evaluates EU and re-balances domain coverage via dynamic sampling budget allocation based on the updated model state θ_t.

EU Estimation The first step in each iteration is to identify documents that bridge the model's current knowledge gap regarding the target domain. Prior EU estimation, such as Entropy, relies solely on model variance, often failing to detect domain misalignment, as illustrated in Figure 2-(c, High Entropy). In contrast, we measure EU by contrasting the model's representation with the target domain statistics.

Prior to the iterative sampling, we pre-compute the target domain statistics, specifically token-level IDF, as illustrated in Figure 2-(b). Then, we project

the document embedding e_d to the vocabulary space via the model's MLM head to obtain token probabilities $p(t|e_d;\theta_t)$, see Figure 2-(c). For the document embedding, we follow the model's pooling method, that is, mostly mean pooling for an encoder-based model, and last token pooling for a decoder-based model (details in subsection A.4. A document is considered informative (high EU) when the model fails to predict high-IDF domain terms, in contrast to low EU documents where predictions are confident. Formally, the EU score $U_k(d)$ aggregates the discrepancy between the domain importance (IDF) and the model's prediction for the top-k tokens $T_k(d;\theta_t)$:

$$U_k(d;\theta_t) = \sum_{t \in T_k(d;\theta_t)} [\log \mathrm{IDF}(t) - p(t|e_d;\theta_t)] \tag{3}$$

We set $k = 1000$, which covers approximately 90% of the cumulative probability mass (about 3% of BERT's 32k vocabulary).

Crucially, as the model adapts to the target domain, $p(t|e_d;\theta_t)$ evolves, dynamically altering the set of high EU documents in each round.

Early Stopping Criteria Once the uncertainty scores are calculated, determining when to stop is crucial to prevent overfitting and save costs. As depicted in Figure 2-(d), we hypothesize that the domain-averaged EU reflects the model's knowledge saturation. We monitor this score using an Exponential Moving Average (EMA) to smooth out fluctuations. The loop terminates when the smoothed EU reaches a plateau (local minimum), or the maximum budget (5k) is exhausted. As validated in our analysis, this unsupervised criterion effectively signals the point of peak retrieval performance.

EU Sampling If the stopping criteria are not met, we proceed to sample documents for adaptation training. While uncertainty sampling targets knowledge gaps, it implies a risk of sampling redundant documents within high-uncertainty regions. To ensure coverage across the domain, we adopt the diversity-driven clustering framework of DUQGen (Chandradevan et al., 2024). However, unlike the baseline, which relies solely on diversity, we balance it with our estimated uncertainty. Specifically, within each semantic cluster $\mathcal{C}_i$, we prioritize documents using a Maximal Marginal Relevance (MMR) approach (Carbonell and Goldstein, 1998). We select the top-n_i documents that maximize the following joint score within each cluster for the current retriever model state θ_t:

$$\mathrm{score}(d;\theta_t) = \lambda\, \widehat{\mathrm{U}_k(d;\theta_t)} + (1-\lambda)\, \widehat{\Psi(d;\theta_t)} \tag{4}$$

where n_i indicates the sampling budget for the i-th cluster, $\widehat{\cdot}$ denotes z-score normalization, and Ψ is the diversity score from DUQGen. This strategy ensures that the final training set $S = \bigcup S_i$ is composed of documents that are both informative (high EU) and representative of diverse topics (high diversity).

Addressing EU Shift via Iteration A critical challenge in domain adaptation is the dynamic nature of EU, which evolves as the model trains. According to active learning theory (Settles, 2009), the EU of documents shifts throughout the training process, gradually decreasing in regions where the model has already been exposed to training samples. Consequently, clusters that initially exhibited high EU progressively lose their informativeness for further adaptation to the target domain. This necessitates an iterative sampling that actively avoids redundant selection from regions exhibiting decreased EU. Such an approach ensures that subsequent sampling iterations continuously target the model's evolving knowledge gaps in the target domain rather than repeatedly sampling from clusters where the model has already trained.

Resampling Penalty for Iterative Sampling Prior approaches like DUQGen, however, allocate the sampling budget n_i in proportion to the static cluster size $|\mathcal{C}_i|$, without accounting for the evolving nature of EU. This static allocation strategy neglects EU shifts across training, resulting in the model overfitting to the dominant clusters while neglecting minority clusters where knowledge gaps persist.

To mitigate this limitation, we introduce a resampling penalty that dynamically redistributes sampling attention toward underrepresented clusters exhibiting high EU (Figure 2-(e)). Specifically, the sampling weight w_i for the i-th cluster is adjusted inversely to its accumulated sample count $\mathcal{P}_i$ over previous iterations:

$$w_i = \frac{|\mathcal{C}_i|}{\mathcal{P}_i + \epsilon}, \quad \text{where} \quad n_i = n \cdot \frac{w_i}{\sum_j w_j}. \tag{5}$$

By penalizing redundant sampling from previously explored dominant clusters, this mechanism progressively shifts sampling weight toward underrepresented minority regions.

4 Experimental Setup

4.1 Implementation Details

AU Filtering We utilize PySerini's prebuilt BM25 index to calculate the lexical k-NN distance for filtering low-density documents. We set the neighbor count $k = 3$ and the z-score threshold $Z_{thr} = 1.5$ to flag and exclude outliers in a domain-adaptive manner.

Iterative Sampling-Training Loop For EU estimation, we compute probabilities using the top-1,000 tokens and set the balance weight $\lambda = 0.5$ to equally weigh uncertainty and diversity. In the iterative loop, we sample 500 documents per iteration up to a maximum of 10 iterations (total 5k budget), matching the baseline's scale. The early stopping mechanism employs an EMA with a smoothing factor $\alpha = 0.4$.

Pseudo-Query Generation We employ Llama3-8B-Instruct (Dubey et al., 2024) with the template shown in Appendix A.8. We generate one query per document using temperature 0.8 and top-p 0.9.

4.2 Datasets and Evaluation Metrics

We evaluate our method on five large-scale BEIR datasets (>100k documents) where effective selection is challenging: TREC-COVID (TC), Robust04 (RB), TREC-NEWS (TN), Quora (QR), and HotpotQA (HQ). Dataset statistics are detailed in Table 7.

To evaluate the retrieval performance of the adapted model, we use normalized Discounted Cumulative Gain (nDCG@10), which measures the number and order of relevant documents ranked in the top-10 retrieved documents.

4.3 Baselines

We compared UnIte with four document selection strategies under a fixed budget of 5k documents:

- Random: the documents are randomly sampled.
- GPL (Wang et al., 2022): the documents are randomly sampled, and the relevance annotations between pseudo query and document are labeled with an expensive cross-encoder. For fair comparison, we adapt it to generate 5 k training samples.
- Quality (Chang et al., 2024): the documents are sampled by a neural quality estimator model.
- DUQGen (Chandradevan et al., 2024): the diversity-based sampling via clustering with external embedding model, Contriever (Izacard et al., 2022).

4.4 Retrieval Models

We assess performance using four single-vector retrievers initially trained on MS-MARCO (Bajaj et al., 2018): DPR (Karpukhin et al., 2020), coCondenser (Gao and Callan, 2021), COCO-DR (Yu et al., 2022), and Qwen3-Embedding-4B (Zhang et al., 2025) [3] All models are fine-tuned on the selected documents using their standard objectives. Note that all experiments were conducted on a single NVIDIA 3090 GPU, and the detailed configurations are described in Appendix A.2.

5 Results

5.1 Overall adaptation performance

Our results in Table 1 show that UnIte consistently improves nDCG@10 across retrievers and document selection methods, while Random yields occasional gains and Quality sometimes degrades performance. On average, UnIte improves nDCG@10 by +2.45, +0.75, and +0.26 points over DUQGen with DPR, coCondenser, and COCO-DR, respectively. For DPR, UnIte outperforms GPL by a large margin, highlighting the importance of uncertainty-based sampling with limited training data. With Qwen3-embedding-4B, gains increase to +3.49 points, demonstrating effective scaling with model capacity (HQ excluded as Qwen3-embedding was trained on it (Zhang et al., 2025)). These results confirm that uncertainty combined with diversity yields the most robust adaptation gains.

5.2 Ablation Study

There are two major components in UnIte that correspond to uncertainty taxonomy: (i) *AU Filtering* and (ii) *EU Sampling*. To empirically prove that both components complementarily contribute to the performance gain, we performed an ablation study in Figure 3. We additionally examine the effect of the resampling penalty, which governs budget allocation within the iterative sampling loop, in Table 2.

[3] Additionally, we experiment with ColBERT (Khattab and Zaharia, 2020) and MonoT5 (Nogueira et al., 2020) (results in Appendix A.4).

Retriever	Adaptation Method	Large Corpus					Total AVG
		TC	RB	QR	TN	HQ	
BM25	—	65.59	40.70	78.9	39.8	60.3	44.49
DPR	— †	33.2	25.2	24.8	16.1	39.1	27.68
	Random	60.81	31.11	74.01	28.66	38.02	46.52
	Quality	61.11	28.25	75.01	23.42	39.35	45.43
	DUQGen	62.75	31.48	75.17	24.05	39.62	46.61
	UnIte	**66.79**♭ ↑4.04	**33.23*** ↑1.75	**75.32** ↑0.15	**29.13*** ↑5.08	**40.82*** ↑1.20	**49.06**♭ ↑2.45
coCondenser	—	67.48	32.51	86.36	28.9	54.44	53.94
	Random	66.35	32.26	87.12	28.14	56.00	53.97
	Quality	63.67	31.36	86.61	**31.95**	56.17	53.95
	DUQGen	70.35	32.42	87.12	28.02	**56.77**	54.94
	UnIte	**71.02** ↑0.67	**33.95*** ↑1.53	**87.17** ↑0.05	31.16* ↑3.14	55.14 ↓1.63	**55.69**♭ ↑0.75
COCO-DR	—	79.34	44.64	86.73	**38.61**	60.43	61.95
	Random	79.18	44.8	87.05	38.34	60.53	61.98
	GPL	78.38	43.72	86.95	37.3	59.82	61.23
	Quality	79.12	45.09	86.88	37.96	60.6	61.93
	DUQGen	79.16	45.05	87.15	37.84	**60.84**	62.01
	UnIte	**80.02*** ↑0.86	**45.29** ↑0.24	**87.16** ↑0.01	38.31♭ ↑0.47	60.56 ↓0.28	**62.27**♭ ↑0.26
Qwen3 (4B)	—	88.91	62.27	88.30	21.17	-	65.16
	GPL	88.81	59.84	88.28	21.89	-	64.71
	DUQGen	88.60	60.90	83.90	43.82	-	69.31
	UnIte	**91.60*** ↑3.00	**62.31*** ↑1.41	**88.62*** ↑4.72	**48.68*** ↑4.86	-	**72.80*** ↑3.49

Table 1: Retrieval performance (nDCG@10) on BEIR across retrievers and adaptation methods. **Bold** entries mark the highest performances per dataset for each retriever, while the underlined entries indicate the second-highest. "AVG" columns report the overall average. † indicates values taken from the original paper. ↑ is the difference between UnIte and DUQGen, and statistically significant improvement over DUQGen is denoted as $*$ ($p < 0.05$) and ♭ ($p < 0.1$).

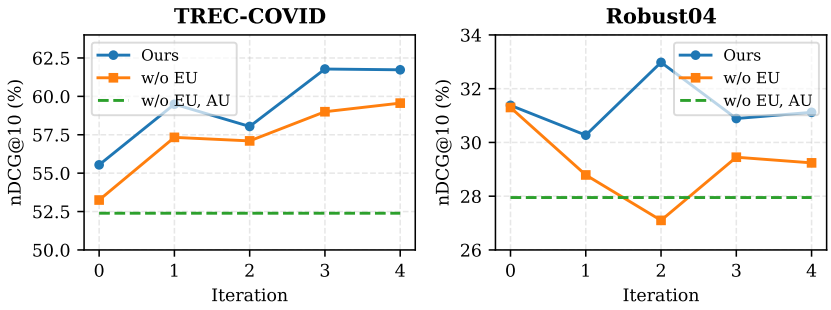

Figure 3: Ablation results in an iterative sampling-training loop with and without AU and EU. Without both stages, we only measure the first-iteration performance and plot it as a horizontal line. We report DPR performance in nDCG@10 on TREC-COVID and Robust04.

Impact of *EU Sampling* First, we compared the performance with and without EU sampling with DPR on TREC-COVID and Robust04. Figure 3 shows that UnIte (blue line) consistently outperforms the baseline without EU sampling (orange line, "w/o EU") across all iterations on both datasets. Notably, on Robust04, the performance without EU sampling drops significantly—by approximately 4 nDCG@10 points—compared to the zero-shot. This performance gap indicates that document sampling without considering the model's understanding leads to redundant training samples.

Impact of *AU Filtering* To further assess the contribution of AU filtering, we conducted an additional experiment by removing this component. Without both modules (green dashed line, "w/o EU, AU"), we sampled all 5k samples at once, as the method cannot iterate without these components. This configuration shows a substantial performance gap of approximately 5 and 9 nDCG@10 points compared to the peak performance of UnIte on TREC-COVID and Robust04, respectively. This demonstrates that documents sampled from low-density regions critically contribute to false posi-

Setting	TC	QR	TN
w/ Resampling Penalty	**61.73**	**74.95**	**30.39**
w/o Resampling Penalty	54.39	73.42	21.83
Δ	+7.34	+1.53	+8.56

Table 2: Impact of resampling penalty Equation 5 on nDCG@10 with DPR training with 2.5k samples.

tives when AU filtering is absent.

Impact of Resampling Penalty To assess the contribution of the resampling penalty Equation (5), we conducted an ablation study using DPR, and for fair comparison, we fixed the sample size to 2.5k. As shown in Table 2, removing the penalty consistently degrades performance across datasets by about 1.5 to 8.5 nDCG@10, confirming that dynamic budget redistribution prevents over-sampling from dominant clusters.

Hence, removing either of the AU and EU modules harms adaptation, and further performance improvement is largely unattainable. Moreover, the resampling penalty complementarily works to ensure balanced coverage across clusters, preventing the iterative loop from oversampling dominant regions.

5.3 EU estimation methods

In prior work, EU has predominantly been discussed solely in terms of the model itself. However, in UDA, it is important to measure the model understanding of *target domain* to properly estimate the EU, by introducing distribution statistics.

To assess the effectiveness of incorporating the target domain distribution into EU estimation for UDA, we evaluated adaptation performance under a fixed sampling strategy while varying the estimation methods. Specifically, we conducted experiments on TC, RB, and TN with DPR, using a budget of 500 documents. Table 4 shows that our EU estimation method using target domain distribution consistently outperforms other estimation methods across all three domains, yielding an average improvement of 2.53 nDCG@10, over MC-Dropout. Since entropy and MC-dropout EU estimation overlook target-distribution signals, their estimations are consequently less domain-aware, which results in degraded retrieval performance after adaptation. Overall, our findings explicitly demonstrate that effective document selection for domain adaptation requires incorporating the target domain distribution into EU estimation, which enables EU to focus on more valuable samples.

5.4 Early Stopping Criteria

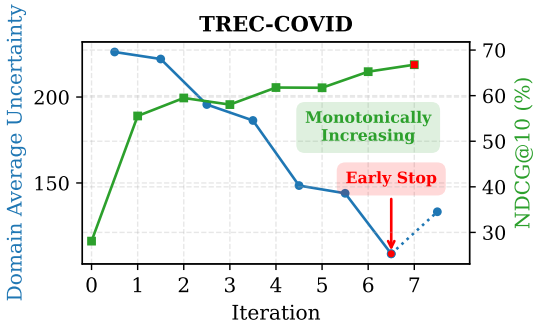

Figure 4: Relationship between model performance and average uncertainty across the iterative sampling-training loop. Average EU across the target domain (left axis) and nDCG@10 (right axis) is reported on TREC-COVID.

While increasing the number of training documents may appear beneficial, prior work has demonstrated that excessive samples can degrade performance through overfitting (Chandradevan et al., 2024). As illustrated in Figure 4, the average EU initially decreases as the model adapts to the target domain, then subsequently increases, indicating sample redundancy. Notably, the local minimum in average EU coincides with peak retrieval performance, validating EU as a reliable early stopping criterion. This characteristic enables UnIte to achieve both effectiveness and sample efficiency. As demonstrated in Table 3, UnIte achieves a 0.94 nDCG@10 improvement per 1k samples with DPR compared to DUQGen, and typically converges at 3–5k samples, in contrast to DUQGen, which utilizes 5k samples.

5.5 Computational Cost

The sample efficiency achieved through early stopping directly reduces computational costs through smaller training datasets and fewer adaptation iterations.

Training time Training DPR with the baseline 5k dataset requires approximately 10 minutes on a single NVIDIA 3090 GPU, whereas ours utilizes early stopping, which typically produces 3-5k samples, completing adaptation in around 8 minutes.

Dataset construction time For the domain adaptation method, most of the additional computation is due to constructing the adaptation dataset. For our method, AU filtering is applied once on the

Retriever	Adaptation Method	Large Corpus					Total AVG
		TC	RB	QR	TN	HQ	
DPR	DUQGen	5.91	1.26	10.07	1.59	0.10	3.56 (5k)
	UnIte	**9.6** (3.5k)	**1.61** (5k)	**10.1** (5k)	**3.26** (4k)	**0.34** (5k)	**4.50** (4.5k)
coCondenser	DUQGen	0.57	-0.02	0.15	-0.18	**0.47**	0.20 (5k)
	UnIte	**1.18** (3k)	**0.29** (5k)	**0.2** (4k)	**0.5** (4.5k)	0.14 (5k)	**0.41** (4.3k)
COCO-DR	DUQGen	-0.04	0.08	0.08	-0.15	**0.08**	0.01 (5k)
	UnIte	**0.14** (5k)	**0.13** (5k)	**0.09** (5k)	**-0.06** (5k)	0.03 (5k)	**0.06** (5k)
Qwen3-Embedding-4B	DUQGen	-0.06	-0.27	-0.88	4.53	-	0.83 (5k)
	UnIte	**0.54** (5k)	**0.01** (5k)	**0.06** (5k)	**9.17** (3k)	-	**2.45** (4.5k)

Table 3: Retrieval performance gain over zero-shot for unit sampling size (Δ nDCG@10 / 1k). Total sampling size is shown in brackets, except for DUQGen, which always samples 5k. **Bold** values indicate better performance per dataset for each retriever.

	TC	RB	TN	AVG
UnIte	**55.54**	**31.38**	**23.33**	**36.75**
MC-Dropout	52.79	28.27	21.6	34.22
Entropy	54.1	25.83	22.7	34.21

Table 4: Comparison of uncertainty measures with UnIte. DPR performance (nDCG@10) on the first iteration is reported on TC, RB, TN, and their average (AVG). The best result for each column is highlighted in **bold** and the second-best result is underlined.

(a) Impact of Z_{thr}

Method	Z_{thr}	nDCG	Ratio
Zeroshot	-	28.58	-
UnIte w/o EU	1.0	28.57	10%
	1.5	**29.27**	5%
	2.0	28.82	3.8%
	2.5	28.82	1.2%
	3.0	27.47	0.4%

(b) Impact of λ

Method	λ	nDCG
DUQGen	-	15.37
UnIte	0.1	15.37
	0.3	14.16
	0.5	**16.32**
	0.7	15.32
	0.9	15.33

Table 5: Hyperparameter tuning. Left: Impact of Z_{thr} (CoCondenser). Right: Impact of λ (DPR).

full corpus $\mathcal{C}$, requiring 120 seconds, and EU estimation operates on the filtered corpus $\mathcal{C}'$ at each iteration, requiring 150 seconds per iteration. Since EU scales linearly with corpus size, the overhead remains practical even for larger corpora. The total overhead is $120 + 150 \times N$ seconds, where N is the number of iterations until convergence. As our method early stops at 4k samples on average, the net time savings are 880 seconds[4]. Even with 5k samples, the overhead remains modest at approximately 20 minutes, demonstrating reasonable computational cost for superior performance.

5.6 Hyperparameter Tuning

We conducted hyperparameter tuning on FiQA, the largest dataset under 100k documents in BEIR, and applied the selected values across all other datasets. Note that tuning directly on target domains would constitute information leakage in our out-of-domain evaluation setting. Therefore, we selected FiQA to serve as a held-out development set. As shown in Table 1, the fixed hyperparameters yield consistent improvements across all five evaluation datasets, supporting their cross-domain robustness.

Filtering Threshold ($\mathcal{Z}_{thr}$) In Equation (2), the threshold $\mathcal{Z}_{thr}$ determines which documents to filter based on AU, to remove approximately 5-10% of documents that are poorly aligned with the target domain. Since our density-based filtering relies on normalized scores, the threshold can be adjusted to control the removal percentage. We evaluated both retrieval performance and document filtering ratios for $\mathcal{Z}_{thr} \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$. As presented in Table 5, using coCondenser, $\mathcal{Z}_{thr} = 1.5$ achieves optimal performance of 29.27 nDCG@10 while filtering approximately 5% of documents. More strict thresholds ($\mathcal{Z}_{thr} = 1.0$) remove an excessive proportion of documents (10%), resulting in degraded performance of 28.57 nDCG@10, whereas more loosened thresholds ($\mathcal{Z}_{thr} \geq 2.0$) retain noisy outliers, leading to performance degradation (nDCG@10 $\leq$ 28.82).

Balance Weight (λ) We set $\lambda = 0.5$ to equally balance uncertainty and diversity in Equation (4). To validate this choice, we conducted a hyperparameter sweep using DPR with $\lambda \in$

[4] $2.2 \times 1000 - 120 - 150 \times 8$

$\{0.1, 0.3, 0.5, 0.7, 0.9\}$. We excluded the boundary values of 0 and 1, as these would entirely neglect one component. As demonstrated in Table 5, $\lambda = 0.5$ yields optimal performance, substantially outperforming extreme values that disproportionately emphasize either uncertainty ($\lambda = 0.1$, nDCG@10=15.37) or diversity ($\lambda = 0.9$, nDCG@10=15.33). The balanced configuration of $\lambda = 0.5$ confirms that neither metric alone is sufficient for effective sampling.

Smoothing Factor (α) For early stopping criteria, we applied Exponential Moving Average (EMA) smoothing factor with $\alpha = 0.4$ to address noticeable fluctuations in the uncertainty measure. With a step size of 10, $\alpha = 0.4$ effectively corresponds to a weighted window of approximately 4 steps. This parameter selection is critical for early stopping: certain datasets terminate after only 2.5k samples (5 steps), and excessive smoothing ($\alpha > 0.5$) over-attenuates the signal, pulling it too strongly toward the global average and obscuring the underlying trend. We selected $\alpha = 0.4$ to balance noise reduction and trend preservation given the limited number of steps. As illustrated in Figure 6, smaller values (e.g., $\alpha = 0.3$) exhibit similar early stopping behavior and do not noticeably affect performance, but demonstrate reduced stability on average compared to $\alpha = 0.4$.

6 Conclusion

We studied document selection for UDA retrievers and found that DUQGen's diversity-only strategy tends to oversample low-density and high-confidence regions. UnIte addresses this by filtering noisy documents via AU and prioritizing informative ones via iterative EU sampling. This model-conditioned approach yields consistent gains across five BEIR corpora and four retrievers with fewer pseudo-queries. Our results highlight uncertainty-aware sampling as a promising direction for budget-efficient domain adaptation.

7 Limitations

While our method consistently improves retrieval performance for single-vector retrievers, our treatment of multi-vector and re-ranking architectures is preliminary. As reported in Appendix A.4, we approximate single-vector behavior by applying pooling and a shared vocabulary projection layer. However, these processes depart from the model's native training objectives, limiting interpretability. Developing protocols that align with each architecture's objective is a direction for future work. Finally, our estimate of the target distribution relies on IDF statistics. Exploring richer distribution measures (e.g., topic-conditioned statistics, document-frequency variants) is another promising direction.

Lastly, our resampling penalty mitigates bias toward dominant clusters but does not explicitly account for rare minority topics that might be relevant to the target domain. In domains with highly skewed topic distributions, such topics may still be underrepresented in the final training set, potentially limiting adaptation on those topics. Evaluating the impact of our sampling strategy on minority topic coverage is an important direction for future work.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00414981), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00077/RS-2022-II220077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *Preprint*, arXiv:2202.05144.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Ramraj Chandradevan, Kaustubh D. Dhole, and Eugene Agichtein. 2024. Duqgen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. *Preprint*, arXiv:2404.02489.

Xuejun Chang, Debabrata Mishra, Craig Macdonald, and Sean MacAvaney. 2024. Neural passage quality estimation for static pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 174–185, New York, NY, USA. Association for Computing Machinery.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Sho-

janazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *Preprint*, arXiv:2108.05540.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query–: When less is more. *Preprint*, arXiv:2301.03266.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. *Preprint*, arXiv:2202.02794.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Contrastive pre-training for zero-shot information retrieval.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Dawn Lawrie, Efsun Kayi, Eugene Yang, James Mayfield, Douglas W. Oard, and Scott Miller. 2025. Generate-distill: Training cross-language ir models with synthetically-generated data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2926–2930, New York, NY, USA. Association for Computing Machinery.

Dohyeon Lee, Jongyoon Kim, Jihyuk Kim, Seung-won Hwang, and Joonsuk Park. 2025. tRAG: Term-level retrieval-augmented generation for domain-adaptive retrieval. In *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)*, pages 6566–6578, Albuquerque, New Mexico. Association for Computational Linguistics.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. *Preprint*, arXiv:2004.14503.

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

Burr Settles. 2009. Active learning literature survey.

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

Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *Preprint*, arXiv:2112.07577.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *Preprint*, arXiv:2210.15212.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Appendix

A.1 Algorithmic Description

Algorithm 1 describes the formal procedure of UnIte. Initially, the corpus is refined by filtering out high AU samples (Line 1). The method then proceeds iteratively through T rounds. In each iteration, we first assess the EU of the remaining candidates and monitor convergence (Lines 4–7). We then perform adaptive sampling, where cluster weights are dynamically adjusted to penalize redundancy, ensuring the selection of a diverse batch S_t (Lines 9–13). Finally, the retriever θ is updated using pseudo-queries generated from S_t (Lines 15–16). This iterative interaction allows the model to progressively adapt to the target domain by focusing on informative yet reliable samples.

A.2 Fine-tuning details

We form a synthetic training set $\{(q_d, d)\}$ for $d \in S$, mining negatives via:

- *In-batch negatives:* other documents $d' \in S$ in the same batch.
- *Contriever hard negatives:* bottom negatives of the top-100 Contriever retrievals on $\mathcal{D} \setminus \{d\}$.

Bi-encoder models (DPR, coCondenser, COCO-DR) use an InfoNCE loss. DPR was fine-tuned with a batch size of 32, learning rate of $2e^{-5}$, AdamW optimizer with weight decay of $1e^{-2}$, eps of $1e^{-8}$, and maximum sequence length of 509. coCondenser was fine-tuned with a batch size of 32, learning rate of $5e^{-6}$, and other training details follow Tevatron (Gao et al., 2022). COCO-DR was fine-tuned with a batch size of 32, learning rate of $1e^{-6}$, and also followed the Tevatron train setup. [5] Fine-tuning runs on NVIDIA 3090 GPU, with each epoch completing in approximately 10 minutes. For training GPL Baseline, we follow the data generation process and only modified the selected number of documents to 5k. Learning rate is modified to match the same setup as mentioned above.

A.3 AU Approximation

To find the optimal number of hyper-parameter k in AU approximation using lexical k-NN distance, we use the elbow (Thorndike, 1953) method in each domain. Figure 5 illustrates the elbow point detected in the TREC-COVID domain. For all 5

[5] Trained using RTX 3090 GPU with 24GB memory

Algorithm 1 UnIte: Uncertainty-based Iterative Document Sampling

Require: Unlabeled Corpus $\mathcal{C}$, Initial Retriever θ_0, Iteration limit T, Batch size B

$\mathcal{C}' \leftarrow$ AU_Filtering($\mathcal{C}$) ▷ Remove high aleatoric uncertainty outliers
Initialize sampled set $\mathcal{P} \leftarrow \emptyset$
for $t = 1$ to T **do**
    // Step 1: EU Estimation
    Compute $U_k(d; \theta_{t-1})$ for all $d \in \mathcal{C}'$ using Eq. (3)
    Calculate domain average EU $\bar{U}^{(t)}$
    if Plateau($\bar{U}^{(t)}$) **then break**
    end if ▷ Early Stopping
    // Step 2: EU Sampling
    for cluster c_j in Clusters **do**
        Update weight w_j based on prior selection $\mathcal{P}_j$ using Eq. (5) ▷ Resampling Penalty
    end for
    $S_t \leftarrow$ Select B docs maximizing Eq. (4) with weights w
    $\mathcal{P} \leftarrow \mathcal{P} \cup S_t$
    // Step 3: Model Update
    Generate pseudo-queries for S_t
    Update retriever $\theta_t \leftarrow$ Train(θ_{t-1}, S_t)
end for
return Updated Retriever θ_t

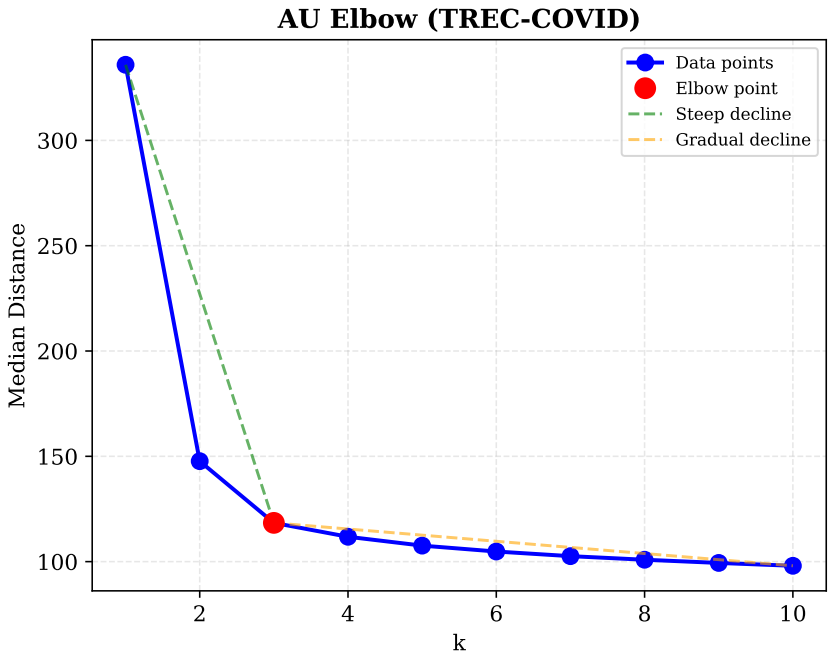

Figure 5: The medians of the lexical kNN distances across various k-values are illustrated.

domains selected in UnIte share the same elbow point as the k-value of 3.

A.4 Other models results

In UnIte method, we assume that there is a single-vector embedding of each document, and apply the MLM Head of the base model to acquire the logit scores. ColBERT is a multi-vector late-interaction model, and MonoT5 reranker is an Encoder-Decoder model. In both cases, we can't naturally derive single-vector embeddings. For the ColBERT model, we drop the token-level embedding projection layer and apply mean-pooling to get the single-vector embeddings. Also, for the MonoT5 model, we take the mean-pooling for the

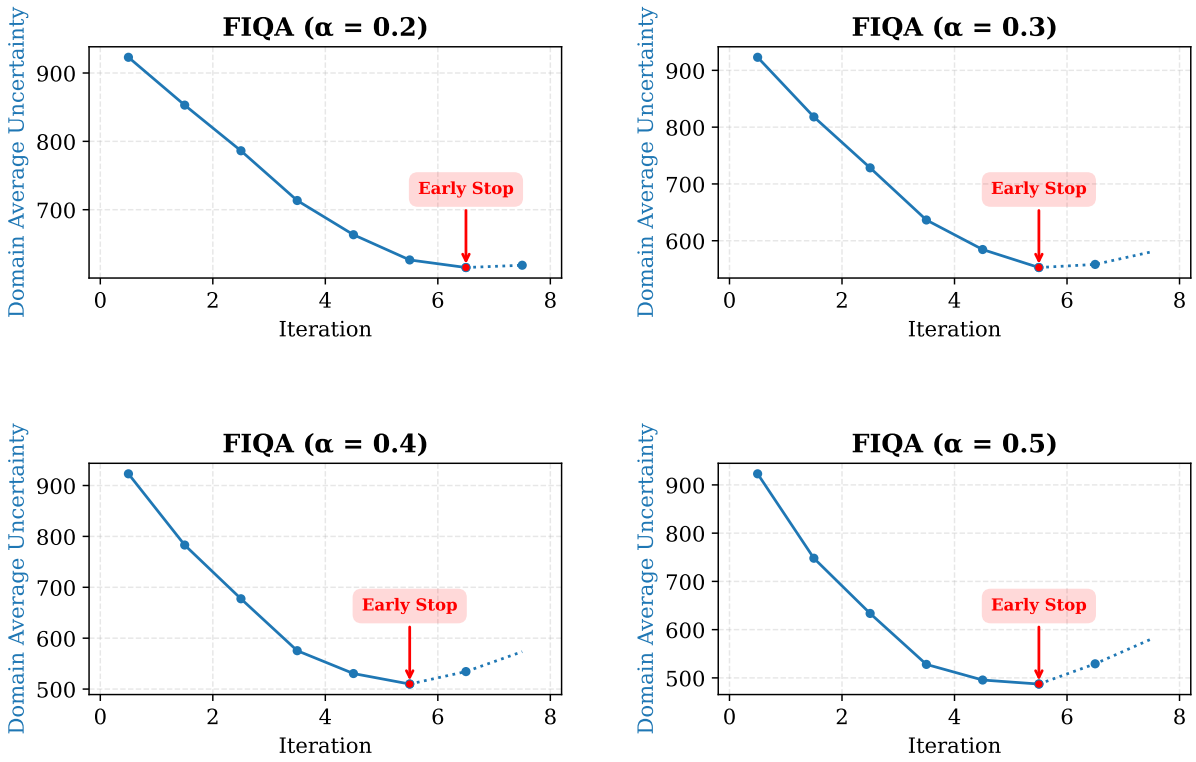

Figure 6: Comparison of smoothed graph trends across different settings of smoothing factors.

Encoder's last hidden state and use it as a single-vector embedding. For the language modeling head, T5 LM Head is used. Since this LM Head is not trained to accept the encoder hidden state, there might be some errors in the underlying logit value. The results are illustrated in Table 6. Although not every domain wins against the baseline, the average performance of UnIte outperforms the baseline in MonoT5.

ColBERT uses a cross-entropy over token interactions. We fine-tuned with official hyperparameters, which are the batch size of 32, the learning rate of $3e^{-6}$, and the maximum sequence length of 300. MonoT5 uses a next-token prediction loss, and fine-tuned MonoT5-base using a batch size of 8, a learning rate of $2e^{-5}$, Adafactor optimizer with weight decay of $5e^{-5}$, and the warp-up ratio of 0.1, regression loss, and maximum sequence length of 512.

A.5 Dataset Statistics

In this paper, we focus on five BEIR benchmark datasets (Thakur et al., 2021): TREC-COVID, Robust04, Quora, TREC-NEWS, and HotpotQA. The statistics for each dataset, including the number of documents, test queries, and relevant documents per query, can be found in Table 7.

A.6 Licenses for Artifacts

All datasets and pretrained models were used strictly for research purposes, in accordance with their intended use and license terms. All datasets are English-only, and no demographic annotations were used. We complied with all license restrictions, including non-commercial clauses where applicable (e.g., DPR under CC BY-NC 4.0). The datasets are publicly available via BEIR on GitHub[6] and Hugging Face[7] under their respective licenses. Pretrained models were obtained from Hugging Face under their original licenses: DPR (CC BY-NC 4.0), coCondenser (Apache-2.0), COCO-DR (MIT), ColBERT (MIT), and MonoT5 (MPL-2.0). We will release our code and training scripts for research and educational use only, which is compatible with the access conditions of the original resources.

A.7 Use of Packages

We used `spaCy` (Honnibal et al., 2020) for text pre-processing before document sampling, includ-

[6] https://github.com/beir-cellar/beir
[7] https://huggingface.co/BeIR

Retriever	Adaptation Method	Large Corpus					Total AVG
		TC	RB	QR	TN	HQ	
First-stage Retriever							
BM25	—	65.59	40.70	78.9	39.8	60.3	44.49
ColBERT	— †	70.6	39.2	85.3	39	59	58.62
	DUQGen	**74.18**	44.95	**85.57**	36.93	**63.44**	**61.01**
	UnIte	73.43	**46.37**	85.09	**37.48**	60.64	60.6
Reranking BM25 Top-100							
monoT5	— †	81.39	51.81	84.65	45.9	69.79	66.71
	InPars	80.3	51	—	—	—	-
	DUQGen	84.76	**54.04**	**88.17**	45.04	**71.54**	68.71
	UnIte	**85.09**	53.91	88.1	**45.79**	71.25	**68.83**

Table 6: Retrieval performance (nDCG@10) on BEIR across late interaction and reranking model and adaptation methods. **Bold** entries mark the highest performances per dataset for each retriever, while the Underlined entries indicate the second-highest. "AVG" columns report the overall average. †indicates values taken from the original paper, and * denotes a statistically significant improvement over DUQGen ($p<0.05$).

	TREC-COVID	Robust04	Quora	TREC-NEWS	HotpotQA
Domain	Bio-Medical	News	Quora	News	Wikipedia
Total # Queries	50	249	10,000	57	7,405
Total # Documents	171.3k	528k	522k	594k	5.2M
Relevant Document / Query	1326.7	1250.6	1.6	53.35	2.1

Table 7: Detailed statistics of the seven subtasks in the BEIR Benchmark that are employed in our experiments. This table includes the number of queries, the number of documents, and the number of relevant documents for a query, for each subtask.

ing tokenization and stop-word removal for inverse document frequency (IDF) computation. We employed `Tevatron` (Gao et al., 2022) to train `coCondenser` and `COCO-DR`.

A.8 Prompts

Example 1:
Document: December 25, 1990, Tuesday, Orange County EditionA mobile-home fire that killed an elderly woman Sunday night was accidental andstarted in her bed, Orange County Fire Department officials said Monday."Some sort of smoking materials in the bedding ignited the fire," said KathleenCha, a County Fire Department spokeswoman.The 75-year-old woman, whose name has been withheld pending notification ofrelatives, . . .
Relevant Query: What caused the fatal mobile-home fire in Dana Point that killed an elderly woman?.
Example 2:
Document: 930818A LABOUR government would impose a levy of up to 1.5 per cent of payrollcosts on companies which failed to comply with training guidelines, MrGordon Brown, shadow chancellor, said yesterday.The levy, intended to help pay for upgrading government training programmes,compares with earlier plans for a maximum levy of 0.5 per cent on allcompanies not spending that amount.The revised proposal emerged in a paper for Labour's annual conference nextmonth, in which Mr Brown further distances the party from thehigher-taxation manifesto on which it fought the 1992 general election.Promising to cut taxes 'if I can', Mr Brown confirmed the Labourleadership's determination to discard the party's redistributionist image.'Labour is not against wealth, nor will we seek to penalise it,' he said.Mr Brown said the revised training proposals were aimed at encouragingcompanies to develop their own training programmes, rather than rely on thegovernment.'There are a large number of companies which are failing to make thetraining investment which is necessary. That is not only harming the countryas a whole, it is harming those companies which are prepared to make theinvestment because they are finding that their trained workers are being ...
Relevant Query: What are Labour's proposed training levy guidelines for companies and the rationale behind them?
Example 3:
Document: BFN[Unattributed report: "A Loose EU Is Not Necessarily to OurAdvantage"][Text] As a member of the European Union [EU], Finlandmust not become subservient to the interests of any of the bigpowers in the EU. Finland must not align itself with either theBritish or the French ideology, but act as defender of theinterests of Europe's northeast corner. A loose EU is notnecessarily advantageous to Finland.These were the points that SDP [Social Democratic Party]Chairman Paavo Lipponen stressed in his speech at a seminarorganized by the Trade and Industry Delegation on Wednesday [4May]. Lipponen pointed out that Finland has eight months toprepare its membership policy and get ready to take fulladvantage of membership. He added that Finland cannot functionfor a single day in the EU without a clearly defined platform ofpolicy."Will it be the French or the British philosophy? The Frenchfocus on finality, a clearly defined goal for a closelyintegrated, federalist EU. The British prefer to advancepragmatically and favor a loose community until this is provenwrong." . . .
Relevant Query: What stance does Finland's SDP Chairman Paavo Lipponen take on a loose versus closely integrated EU, and how does this relate to Finland's interests?
Example:
Document: {document_text}
Relevant Query:

Figure 7: Prompt template with in-context examples for query generation for the Robust04 Dataset.

A.9 Use of AI Assistants

We used ChatGPT for grammatical corrections.