DADA: Distribution-Aware Domain Adaptation of PLMs for Information Retrieval

Dohyeon Lee[1]*, **Jongyoon Kim**[2*], **Seung-won Hwang**[1]†, **Joonsuk Park**[3,4,5†]
Seoul National University[1],
Interdisciplinary Program in Artificial Intelligence, Seoul National University[2],
NAVER AI Lab[3], NAVER Cloud[4], University of Richmond[5]
{waylight3, john.jongyoon.kim, seungwonh}@snu.ac.kr
park@joonsuk.org

Abstract

Pre-trained language models (PLMs) exhibit promising retrieval performance in various domains. However, they struggle in domains unseen during training, since the word distribution can shift significantly. To remedy this, GPL, a generative domain adaptation (DA) method, was proposed to generate pseudo queries and labels for documents in unseen domains to further train the retriever model. However, the pseudo queries often do not resemble real queries from the target domains, as they do not integrate the domain's distributional information. we propose **Distribution-Aware Domain Adaptation** (DADA) to guide the model to incorporate the term distributions at both the document-level and the corpus-level, which we refer to as observation-level and domain-level feedback, respectively. Empirical results on five distinct datasets demonstrate that our method effectively adapts the model to target domains and expands document representation to unseen gold query terms. [1]

1 Introduction

Recent advances in pretrained language models (PLMs) (Devlin et al., 2018; Clark et al., 2020) have significantly enhanced our ability to retrieve information (Yates et al., 2021; Nogueira and Cho, 2019). These advancements are particularly pronounced when the data closely aligns with what the PLM was originally trained on, denoted as "in-domain" data. However, challenges emerge when dealing with "out-of-domain" (OOD) data, characterized by substantial disparities from the training data. Studies such as Thakur et al. (2021) have highlighted these challenges.

In response to these issues, various domain adaptation (DA) methods have been proposed. A widely employed technique is Generative Pseudo-query Learning (GPL) (Wang et al., 2021), which entails generating pseudo queries linked to specific documents to enhance PLM learning. However, the efficacy of these methods hinges heavily on the quality of these pseudo queries. Poorly constructed queries can significantly diminish the effectiveness of the method.

Method	Information Level		Update	
	Observation	Domain	Data	Loss
GPL	✓		✓	
DADA (Ours)	✓	✓		✓

Table 1: Taxonomy based on the information level and update mechanism to classify the baseline method (GPL) and our approach (DADA), following Chen et al. (2023). The information level indicates whether the method utilizes information specific to a particular data point (observation-level) or encompasses broader data trends (domain-level). The update mechanism illustrates how this information is utilized, whether it involves updating the training data or the loss function.

DA techniques can be categorized by two main axes, as illustrated in Table 1, based on Chen et al. (2023). The first category examines the level of information used by DA methods, either focusing on individual document distribution (observation-level) or broader corpus distribution (domain-level). The second category considers how these methods integrate such feedback, either by updating the training data or the loss function used for learning.

DA methods, like GPL, aim to generate pseudo queries that approximate "gold query terms" from users at test time. However, GPL fails to generate terms that are unseen from in-domain training data, when failing to capture the following signals:

- Expanded document observation: Gold query terms may not appear in the observed document, but it might be in its expanded vocabulary including synonyms of the document.

*Both authors contributed equally to this research.
†Corresponding Authors
[1]The code and experimental details are available at https://github.com/ldilab/dada.

Findings of the Association for Computational Linguistics ACL 2024, pages 13882–13893
August 11-16, 2024

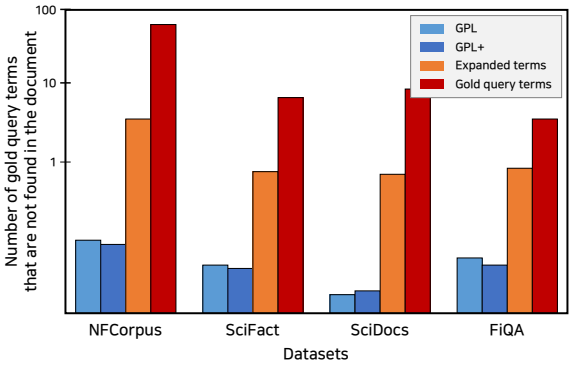

Figure 1: The number of gold query terms that are not found in the document. GPL injects the query q generated from the document d as the target, GPL$^+$ uses q generated from d and the expanded terms.[2]

- Contrasted domain distribution: Gold query terms may be infrequently seen in the training corpus but more frequent in the target corpus, as approximated from its subset.

Table 1 demonstrates that GPL concentrates on observation-level feedback by providing expanded terms through query generation. However, Figure 1 reveals that the query generated by GPL (orange) includes only a subset of the gold terms that observation-level can provide (light blue). Even when leveraging the expanded terms from observation, such as synonyms, during the query generation process (blue), denoted as GPL$^+$, the number of generated gold terms does not exhibit a significant increase. In essence, the query generation process results in a lack of crucial input at the observation level, diminishing the effectiveness of data updating methods. This limitation motivates the importance of implementing a loss updating methodology.

In contrast, our proposed DA method: **Distribution-Aware Domain Adaptation** (DADA), uniquely incorporates both observation- and domain-level feedbacks. While GPL leans more towards observation-level feedbacks, DADA widens the observation-level feedback and even integrates domain-level feedback, offering a broader statistical perspective. The observation-level feedback is widened by domain-level feedback that can approximate the red bar on Figure 1. To obtain such feedback without access to entire OOD corpus, as some documents are accessible in a real-world scenario, we utilize the information obtained from the subset of OOD corpus. Consequently, DADA prevents the loss of essential information during query generation, enhancing the efficiency and effectiveness of the DA process, as demonstrated in experiments on five corpora from the BEIR benchmark(Thakur et al., 2021).

[2]The analysis of discrepant tendency in SciDocs compare to other 3 datasets can be found on Appendix A.2

2 Related Work

This section delves into the landscape of PLMs within information retrieval (IR) and examines the evolution of DA techniques that have paved the way for our proposed method.

2.1 PLMs in IR

The integration of PLMs, exemplified by BERT (Devlin et al., 2018) and ELECTRA (Clark et al., 2020), has revolutionized IR. These models excel in understanding and processing complex linguistic patterns, significantly enhancing document retrieval and ranking processes. Despite their advancements, challenges arise when applying these models to OOD data, an issue highlighted in studies such as Thakur et al. (2021). This limitation has promoted research into effective DA techniques to ensure the robustness of PLMs across diverse domains.

2.2 DA for PLMs

Addressing the domain shift challenge has led to the development of various DA strategies. A prominent approach involves creating pseudo queries (Ma et al., 2021; Liang et al., 2020) to bridge the gap between the PLM's training domain and new target domains. This strategy is crucial in ensuring the relevance and applicability of PLMs to diverse datasets. GPL (Wang et al., 2021) is a noteworthy example, utilizing a cross-encoder mechanism to enhance the alignment between pseudo queries and target documents. However, these techniques tend to focus on observation-level feedback, often overlooking broader domain-level insights.

2.3 Our Distinction

Our method marks a significant departure from traditional approaches. It uniquely integrates both observation-level and domain-level feedback, with a focus on updating the loss function. This dual-level feedback approach ensures comprehensive DA, overcoming the limitations of previous methods that mainly rely on observation-level data up-

dates. DADA's innovative strategy enhances adaptability and accuracy in handling OOD data, setting a new precedent in DA techniques.

3 Proposed Method

In this section, we provide a detailed explanation of the methods employed in our study, focusing on the integration of the DADA into the GPL framework. We begin with a concise overview of GPL, followed by an in-depth description of how DADA is integrated into GPL, covering both training and inference perspectives.

3.1 Overview of GPL Training Process

Training process of the GPL framework is visually represented in the top of Figure 2. The GPL framework operates by calculating relevance scores between queries and documents using PLMs. Specifically, given a query q and a document d, the relevance score $S(q, d)$ is computed as the dot product of their embedding vectors:

$$S(q, d; M) = M(q) \cdot M(d) \tag{1}$$

where M denotes the PLM.

To differentiate between positive (d^+) and negative (d^-) documents, GPL employs a Margin MSE loss, $\delta(q, d^+, d^-; M)$, defined as the difference between the relevance scores of each document:

$$\delta(q, d^+, d^-; M) = S(q, d^+; M) - S(q, d^-; M) \tag{2}$$

Additionally, GPL utilizes a cross-encoder[3] to predict the relevance scores, resulting in the margin value $\hat{\delta}(q, d)$. The loss function of GPL is to minimize the loss function:

$$\mathcal{L}_{\text{GPL}}(q, d; M) = |\hat{\delta}(q, d) - \delta(q, d; M)|^2 \tag{3}$$

Our goal is to utilize multi-level feedback to update this loss function for more effective DA.

3.2 Integration of DADA into GPL

To enhance DA within GPL, we introduce the DADA method, which incorporates multi-level feedback and loss function update.

3.2.1 Multi-level Feedback Vector

To capture observation-level and domain-level feedback, each feedback is encoded as a vector. The observation-level distribution encompasses terms appearing in a document d and augmented terms, while the domain-level distribution approximates the entire set of documents in the OOD corpus.

Observation-level Distribution As illustrated in blue in Figure 2, observation-level feedback is a process where we take information from a document, d, and turn it into a vector form. This is done using SPLADE [4], a method that allows us to expand the document's original vocabulary, V_d, into an augmented set, $\hat{V}_d$. Once we have this expanded vocabulary, we then create observation-level feedback, R_{obs} as follow:

$$R_{\text{obs}}(d) = [\text{SP}(d, v_1), \text{SP}(d, v_2), \ldots, \text{SP}(d, v_n)] \tag{4}$$

where v_i and SP denote the i-th term of $\hat{V}_d$ and corresponding weight from SPLADE.

Domain-level Distribution The GPL emphasizes terms that are often not important in the given document, hence having a low likelihood of being included in the gold terms. This is due to the training approach that mostly treats most documents as non-relevant. To overcome this limitation, we provide domain-level feedback. The red lines and shapes of Figure 2 shows the key role of domain-level feedback. Domain-level feedback approximate the entire set of documents in the OOD corpus and is not limited to a specific document d. To represent the distribution R_{dom} of these global terms, we employ Inverse Document Frequency (IDF) in this work, as follows:

$$R_{\text{dom}} = [\text{IDF}(v_1), \text{IDF}(v_2), \ldots, \text{IDF}(v_n)] \tag{5}$$

where v_i denotes a term originated from the given OOD corpus and n denotes the number of terms in the corpus. Note that we use up to 1,000 of randomly selected documents to compute the IDF since we assume the oracle IDF distribution is unknown. Just as the use of SPLADE as observation-level distribution can be straightforwardly changed, IDF can be replaced by other domain-level statistics, and key difference between the two that IDF score does not depend on observation d.

Normalization One important consideration when using both observation-level and domain-level distributions together is that their scales can

[3]The cross-encoder parameters are frozen, which is not trainable.

[4]We selected SPLADE in this work, based on its strong performance in predicting unseen query terms in empirical results, but can straightforwardly support other observation-specific statistics, such as DeepImpact (Mallia et al., 2021).

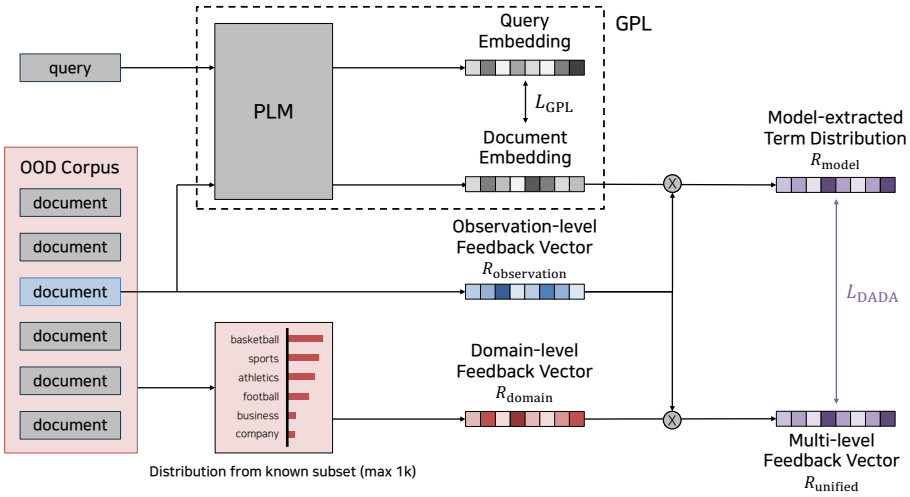

Figure 2: Diagram depicting the overall structure of DADA. The red lines and shapes represent domain-level information which is approximated from known subset (randomly selected maximum 1,000 documents), while the blue lines and shapes represent observation-level (or document-level) information. These two pieces of information are combined in the form of term distributions (purple) and injected into the language model encoder as a loss function.

vary. The different scales of distribution, exceeding the range of 0 to 1, can assign excessively high weights to some specific terms, which will unstabilize the model training. To ensure scale consistency between the two distributions, we first standardize the distribution with z-score normalization, denoted as $Z(\cdot)$, then additionally normalized with tanh, to ensure each term within a single distribution has a weight value between 0 and 1. This two-step normalization can be stated as $\text{Norm}(\cdot) = \tanh\left(Z(\cdot)\right)$.

Merging Two Distributions Finally, as shown in purple in Figure 2, the two distributions R_{domain} and $R_{\text{observation}}$ are integrated into a merged distribution R_{unified} through dot product as follows:

$$R_{\text{unified}}(d) = \text{Norm}(R_{\text{dom}}) \cdot \text{Norm}(R_{\text{obs}}(d)) \quad (6)$$

3.2.2 Loss function update

DADA updates the loss function by injecting combined distribution R_{unified} into GPL. This process entails extracting the term distribution from the model and incorporating it back, thereby aligning the model's distribution with that of the data. The term distribution R_{model} represents what the model has learned and serves as the foundation for updating the loss function. Hence, updating the loss function involves two steps: extracting the term distribution from the model and injecting this distribution information back into the model.

Extraction from model One crucial consideration when extracting term distribution from a model is addressing the discrepancy between embedding space and data space. Specifically, the representations generated by the model reside within the embedding space, whereas the representations obtained from the data exist within the vocabulary space. To overcome this mismatch, we merge these two representations into a single one while respecting the information they both provide.

Taking inspiration from the Max-Sim operation of ColBERT (Khattab and Zaharia, 2020), we can compute the relevance score for each vocabulary term by selecting the maximum value from the dot-product between the embedding vector, generated by the model, and the vocabulary vector obtained from the intermediate output of the observation-level distribution.

$$R_{\text{model}}(d; M) = \max_{i \in [|E|]} (E_i^T \cdot S) \quad (7)$$

$$E = M(d), \quad S = SP(d) \quad (8)$$

where E is the embedding matrix generated by the model M, and S is the vocabulary matrix from the intermediate output of the observation-level distribution.

Injection into model Incorporating data-driven distribution into the model involves reducing the difference between the two distributions. While data distribution interprets the importance of each

vocabulary term, considering both observation-level and domain-level distributions, the model's distribution highlights the relationship between the embedding and vocabulary. The highest value in the vocabulary direction shows the terms the model finds crucial. By minimizing the differences between these two distributions, the model learns the significance of each vocabulary term. Each distribution is normalized with softmax, denoted as $\overline{R}(d)$, to ensure that the distribution represents probabilities.

The final loss funciton of DADA can be defined as:

$$\mathcal{L}_{\text{DADA}}(d; M) = D_{\text{KL}}(\overline{R}_{\text{unified}}(d) \parallel \overline{R}_{\text{model}}(d; M)) \tag{9}$$

where M states the model and D_{KL} means the Kullback-Leibler (KL) divergence (Shlens, 2014).

Therefore, the overall training objective of GPL + DADA is minimizing $\mathcal{L}(d; M)$ which is:

$$\mathcal{L}(d; M) = \mathcal{L}_{\text{GPL}}(q, d; M) + \mathcal{L}_{\text{DADA}}(d; M) \tag{10}$$

where q, d, and M state the query, document, and model, respectively.

3.3 Technique for Stable Training

Merging two distributions can be unstable, as some documents align closely, while others diverge significantly. Therefore, it is important to introduce data in a specific sequence to prevent slow learning. To address this, we use a curriculum that guides the model from simple to complex concepts, enhancing convergence and adaptability to the new corpus. Inspired by curriculum learning (Bengio et al., 2009), we gauge learning difficulty as follows:

$$D_{\text{KL}}(\overline{R}_{\text{dom}} \parallel \overline{R}_{\text{obs}}(d)) \tag{11}$$

Gradually introducing increasingly challenging document distributions during training optimally improves performance, aligning with the concept of the model learning progressively from simpler to more complex data.

3.4 Inference Process of DADA

DADA impacts the training loss of GPL, leaving the inference process unchanged. DADA offers the advantage of enhancing DA without increasing computational costs during inference, ensuring efficient and unaltered operation in this phase. Thus, DADA facilitates optimized DA without compromising the efficiency of inference tasks.

4 Experiments

4.1 Experimental Setup

Implementation Detail In our experiments, we employed two types of feedbacks: SPLADE (Formal et al., 2021) for observation-level and IDF [5] for domain-level feedback.

Given that the target task involves DA, the model was initially trained on the MS-MARCO, as an in-domain dataset. Input document construction involves concatenating document titles and bodies, and truncating sequences longer than 256 tokens. Queries are truncated to a length of 64 tokens.

The representation of query (q) and document (d) are pooled from the output of the model as it is designed. The relevant document is retrieved using the dot product score, $S = E_q \cdot E_d$. Training was performed on four RTX 3090 GPUs with DDP setting, and any unspecified details in the paper follow the same settings as GPL.

Datasets and Evaluation Metrics Our method was evaluated using 5 datasets from the BEIR benchmark (Thakur et al., 2021). Among these, 4 out of 5 datasets (SciFact, SciDocs, FiQA, and NFCorpus) were chosen due to their smaller corpus sizes in the BEIR benchmark datasets, as we focus on scenarios requiring DA, which correlate to limited corpus size. Meanwhile, to observe whether our observation is consistent in a large corpus retrieval scenario, we contrast with Robust04, where the corpus is large and relevant documents can match a query. Detailed information on our selected subtasks is presented in Appendix A.1.

To adapt the model to the target domain, we employed training datasets pairing OOD documents with the pseudo queries and pseudo labels generated by the GPL (Wang et al., 2021) and GPL$^+$.

The evaluation utilized the nDCG@10 metric, widely accepted for assessing the general quality of predictions on the top-10 retrieved documents.

Baselines GPL (Wang et al., 2021) generates pseudo queries and labels using the DocT5Query (Nogueira et al., 2019) query generator and retrievers, incorporating cross-encoders. Both the query generator and retrievers are pretrained on the MS-MARCO dataset (Bajaj et al., 2016). GPL$^+$ closely resembles GPL, with the distinction of incorporating the observation-level distribution into the query generator. The resulting

[5] We use pyserini library to compute IDF.

pseudo queries, generated with this awareness of the observation-level distribution, are subsequently utilized for hard negative mining and pseudo labeling, following the same procedural steps as GPL.

Retrievers (PLMs) To show the robustness of applying DADA along with GPL, we have chosen three different models: coCondenser (Gao and Callan, 2022)[6], COCO-DR (Yu et al., 2022)[7] and TAS-B (Hofstätter et al., 2021)[8]. All three models were trained on the in-domain dataset, MS-MARCO. The models are chosen to show the robust performance across different sizes of model parameters[9].

4.2 Results and Analysis

Research Questions To further validate the effectiveness of our proposed method, we assess whether our goals have been achieved through the following research questions.

- **RQ1**: Is domain-level feedback reflected as intended?
- **RQ2**: Does DADA generate more gold terms as intended?
- **RQ3**: Does DADA adapt better to a new domain as intended?

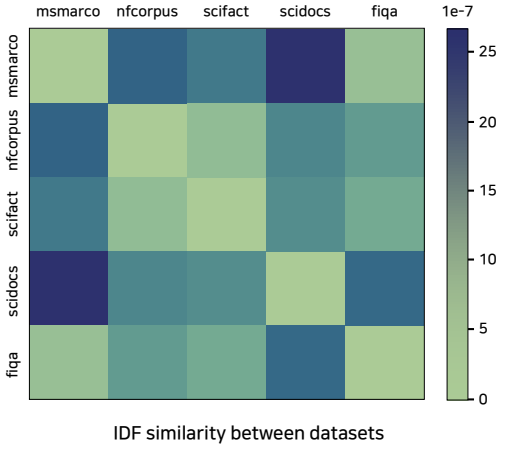

Figure 3: The IDF similarity between datasets is measured using JS-divergence. The scale of each unit is 1e-7. Brighter areas have smaller numbers, indicating that the two IDF distributions are more similar.

[6]Luyu/co-condenser-marco

[7]OpenMatch/cocodr-base-msmarco

[8]sentence-transformers/msmarco-distilbert-base-tas-b

[9]coCondenser and COCO-DR have 110 million parameters while TAS-B has 67 million parameters.

4.3 RQ1: Is domain-level feedback reflected as intended?

To assess the effective integration of domain-level feedback into the model, we employ the transformation technique introduced by Ram et al. (2022), which extracts vocabulary distribution vector v from the model.

This procedure utilizes the Masked Language Modeling (MLM) head of the model to carry out the projection. More precisely, an embedding vector e of a document is subjected to transformation by the MLM Head, leading to its projection onto a vector v with dimensions according to the vocabulary size.

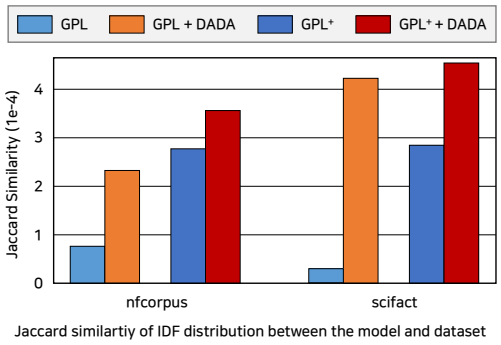

Figure 4: Jaccard Similarity of top-1,000 high-IDF terms distribution between the model and datasets.

Figure 3 illustrates the difference of model-extracted distributions from the IDF distribution of new domain (oracle, which cannot be observed during training), measured using Jensen-Shannon divergence (Nielsen, 2020). Notably, SciDocs, which shows the greatest domain difference from in-domain dataset MS-MARCO, exhibits the lowest GPL performance. Our goal is to adapt the model to perform well in OOD datasets by minimizing these differences in IDF distributions.

To verify whether DADA minimizes the disparities between the IDF distributions depicted in Figure 3, we compare the model-extracted distribution, with the oracle IDF distribution, focusing on the top 1,000 terms with high IDF values.

Figure 4 illustrates the Jaccard similarity (Jaccard, 1912) between the two IDF distributions [10]. The results reveal that DADA demonstrates greater similarity than the baseline models in both the NFCorpus and SciFact datasets, with an increase of

[10]Jaccard similarity quantifies the resemblance between two distributions by calculating the intersection divided by the union, where a higher number of overlapping terms between the two distributions indicates a greater level of similarity

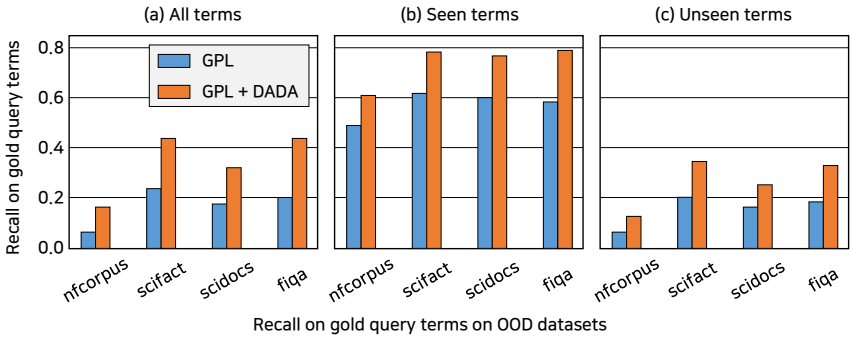

Figure 5: The recall on gold query terms of baseline method (GPL) and our method (DADA). The term "seen" refers to terms from the gold query phrases that appear in the document, whereas the term "unseen" refers to terms that are not appear in the document. A higher recall indicates that the model contains more gold query terms in its embedding space.

more than two times.

However, GPL exhibits lower overall similarity since it does not incorporate domain-level feedback. GPL$^+$ demonstrates greater similarity compared to GPL due to the inclusion an observation-expanded set of terms that includes more relevant phrases. In terms of similarity, the comparison between GPL and GPL$^+$ indicates that the incorporation of domain-level feedback significantly enhances the similarity of IDF distributions.

4.4 RQ2: Does DADA generate more gold terms as intended?

To evaluate whether DADA generates more gold terms, we compare the term distribution from the model (as done in RQ1) and the gold terms query.

Figure 5 illustrates the gold term recall of the top 1,000 terms generated by DADA. Significantly, DADA exhibits superior recall on all OOD datasets, with an improvement ranging from 10% to 20% compared to GPL. This suggests that DADA is more proficient in identifying a larger collection of gold query terms.

Moreover, as depicted in Figure 5(c), the enhanced recall is not solely dependent on identifying seen terms, which are query terms already included in the document. To identify unseen but valuable query terms, both observation-level and domain-level feedback are utilized in DADA, to demonstrate improved ability to identify both seen and unseen terms in comparison to GPL.

For seen terms, DADA not only includes more gold query terms than GPL but also exhibits a similarity of approximately 80% with the oracle IDF distribution. This suggests that using IDF distribution as domain-specific feedback is helpful in determining the weight for seen terms.

4.5 RQ3: Does DADA adapt better as intended?

In order to evaluate how effectively DADA performs in adapting to different domains, we evaluated its performance on five specific subsets of the BEIR benchmark dataset. Table 2 shows the nDCG@10 scores of GPL and DADA for each subset and each retriever. DADA demonstrates significant improvements in the BEIR subset, presenting a distinct advantage over both GPL and GPL$^+$, on various retrievers.

When applied to GPL, DADA enhances retrieval performance in most subsets, resulting in an average nDCG@10 gain of 0.5 in coCondenser, 0.6 in COCO-DR, and 0.1 in TAS-B. Compared to the other two retrievers, where the model size may affect performance, TAS-B exhibits a relatively small performance gain overall and experiences a slight drop in performance on SciFact and FiQA. Since our approach is intended to incorporate additional distribution information through updates to the loss function, smaller models like TAS-B may encounter challenges in integrating this additional information into their parameter updates. However, despite this challenge, there is still an overall improvement.

While DADA also yields performance improvements on coCondenser with GPL$^+$, there is a slight drop in performance on a few datasets. This discrepancy is attributed to the bias in GPL$^+$ construction, focusing too much on expanding document-specific synonyms while missing gold query terms, as also illustrated in Figure 1. FiQA and NFCorpus are such datasets, where the number of unseen

Retriever	Method	SF	SD	FQ	NF	RB	Avg.
coCondenser (Gao and Callan, 2022)	GPL	67.5	**17.1**	32.8	33.7	42.7	38.8
	Ours: GPL + DADA	**67.9**	16.4	**34.1**	**35.1**	**44.7**	**39.6**
COCO-DR (Yu et al., 2022)	GPL	69.7	17.1	33.9	34.4	42.2	39.5
	Ours: GPL + DADA	**69.8**	**17.1**	**35.6**	**34.7**	**42.9**	**40.0**
TAS-B (Hofstätter et al., 2021)	GPL	**67.0**	16.1	**32.9**	33.9	40.4	**38.1**
	Ours: GPL + DADA	66.5	**16.6**	32.2	**34.1**	**40.5**	38.0
coCondenser (Gao and Callan, 2022)	GPL$^+$	63.2	16.2	**32.2**	**33.7**	39.9	37.0
	Ours: GPL$^+$ + DADA	**64.1**	**17.0**	31.8	33.4	**41.2**	**37.5**

Table 2: Comparative evaluation of nDCG@10 scores across different datasets. Experimental settings and parameter configurations used for each algorithm are described in Section 4. The best performance on each dataset for each retriever is highlighted in **bold**. (SF: SciFact, SD: SCIDOCS, FQ: FiQA, NF: NFCorpus, RB: Robust04)

Method	SF	NF	FQ	SD
GPL + DADA	**67.9**	**35.1**	**34.1**	16.4
– R_{domain}	67.3	34.2	33.3	**17.0**
– CL	66.3	34.3	33.0	16.1

Table 3: An ablation study to assess the effects of two new elements introduced by our method (DADA) added to the baseline (GPL): domain-level feedback (R_{domain}) and curriculum learning (CL). For the retriever, coCondenser is used. The best-performing result in each dataset is indicated in **bold**. (Abbreviation of each dataset follows Table 2 notation.)

SF	NF	FQ	SD
12.55	12.60	8.84	6.76

Table 4: Entropy of term distribution of tokenized document in each dataset. All term counts are augmented by 1, to prevent zero logarithm. (Abbreviation of each dataset follows Table 2 notation.)

query terms in GPL$^+$ is fewer than in GPL, where the adverse effect of GPL$^+$ bias overshadows the positive gains from DADA.

We additionally performed an ablation study to assess the influence of two novel components introduced by DADA into the GPL framework: domain-level feedback and curriculum learning. The results, presented in Table 3, showcase the performance of DADA on the SciFact, NFCorpus, FiQA and SciDocs datasets as each of these elements is selectively excluded.

The observed decrease in performance when both elements are omitted serves as compelling evidence of their beneficial impact on model training. This analysis offers valuable insights into the individual contributions of these elements and their collective effect on enhancing the overall efficiency of the model.

Unlike other datasets' performance tendency, SciDocs shows the best performance when the R_{domain} is omitted, which may caused by the nearly uniform R_{domain}, which makes model suffers learning which terms are important in the target domain. The SciDocs corpus includes scientific articles and papers from various fields, which means that various unique words are shown on the corpus unlike other datasets, this can be rephrased as flat term frequency distribution. Measuring the entropy of term frequency of tokenized document in each dataset, as illustrated in Table 4, SciDocs shows the lowest entropy which ensures the nearly uniform distribution of term frequency.

5 Conclusion

We have introduced a novel method called DADA with the specific aim of enhancing the capabilities of PLMs for addressing domain shifts in information retrieval. Unlike traditional methods that resolve domain shifts through dataset updates based on observation-level feedback, DADA takes a direct approach by updating the loss function with a multi-level feedback vector, which integrates domain-level feedback and observation-level feedback. This innovative approach enables DADA to dynamically adjust to varying data landscapes, making it particularly effective in scenarios where rapid adaptation is essential. Compared to an existing method like GPL on the BEIR benchmark, DADA demonstrates superior performance across various subtasks, highlighting its potential in handling unseen domains. While DADA shows promise, further research could investigate integration with other dataset update adaptation techniques, contributing to the development of more robust and versatile information retrieval systems.

6 Limitations

We use DocT5Query to generate queries in experimental setups following the GPL to align with the baseline method. As this query generator has a limitation in effectively generating OOD terms, though we provide domain-level distribution, we overcome this limitation, by directly injecting the domain-level distribution into the PLM. Another possible approach wouldbe upgrading the query to generate OOD terms effectively when such distributions are provided, potentially by utilizing PLMs. We leave this as a future work but expect similar enhancements due to the orthogonal contributions our method offers to GPL.

Acknowledgements

This work has been financially supported by SNU-NAVER Hyperscale AI Center. This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00077, 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.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Valerie Chen, Umang Bhatt, Hoda Heidari, Adrian Weller, and Ameet Talwalkar. 2023. Perspectives on incorporating expert feedback into model updates. *Patterns*, 4(7).

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. 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*, pages 113–122.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1723–1727.

Frank Nielsen. 2020. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221.

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

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to doctttttquery. *Online preprint*, 6.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2022. What are you token about? dense retrieval as distributions over the vocabulary. *arXiv preprint arXiv:2212.10380*.

Jonathon Shlens. 2014. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*.

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

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

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. *arXiv preprint arXiv:2210.15212*.

A Appendix

A.1 Target Datasets

In this paper, we aim to address our experiment on 5 BEIR benchmark datasets (Thakur et al., 2021): SciFact (Lo et al., 2019), SciDocs, FiQA (Maia et al., 2018), NFCorpus, and Robust04. The statistics, including domain information, total number of queries, total number of documents, average query length, average document length, and number of relevant documents for a query, of each dataset can be found on Table 5.

The main focus of our method is improving IR performance in circumstances of small corpus size, where it is difficult to adapt the model to the target domain. For this reason, we initially target 4 datasets that have a smaller corpus size, less than 60k documents: SciFact, SciDocs, FiQA, and NFCorpus. In addition to the 4 datasets, we also included 1 large corpus dataset, Robust04, which has more than 60k documents and many relevant documents that match a query, to demonstrate that DADA can also achieve improvements in a large corpus.

A.2 Discrepant Tendency of Unseen Gold Query Term Comparing GPL and GPL$^+$

The SciDocs dataset, as depicted in Figure 1, contains a slightly larger number of gold query terms, that are not present within the corpus, in GPL (light blue bar) compare to GPL$^+$ (blue bar). GPL$^+$ is designed to include expanded terms during the query generation process, expecting it to capture a wider range of possible queries, which will retrieve more relevant information. However, as the query generator works as a proxy for incorporating the expanded terms, it can neglect or dilute the importance of the terms. In addition, because the other three datasets' corpus term frequency distribution is slightly skewed, which can be found as a high entropy of term frequency in Table 4, it can be more statistically reasonable for the query generator to select seen terms from the given document. For this reason, in most cases, such as NFCorpus, Scifact, and FiQA, the GPL$^+$ shows that the number of unseen gold query terms decreases compared to GPL. Besides the three datasets showing the decrement, the SciDocs shows an increment as the distribution is near uniform, expanding some unseen term can easily increase the value. This marginal improvement in unseen gold query terms, however, does not result in a retrieval performance improvement, in practice, as opposed to the expectation.

A.3 Experiment Results on Additional Metrics

The experiment results are evaluated by various metrics that are often used to evaluate information retrieval performance, such as MRR@10, MRR@100, and nDCG@100. The evaluation results can be found on Table 6, Table 7, and Table 8. The abbreviation of datasets is given as follows. SF: SciFact, SD: SCIDOCS, FQ: FiQA, NF: NFCorpus, RB: Robust04, TC: TREC-COVID)

A.4 Experiment with different observation feedback

The main experiment result shows the impact of using *naver/splade_v2_distil* for observation feed-

	SciFact	SciDocs	FiQA	NFCorpus	Robust04	TREC-COVID
Domain	Scientific	Scientific	Financial	Bio-Medical	News	Bio-Medical
Total # Queries	300	1000	648	323	249	50
Total # Documents	5.2k	25.7k	57.6k	3.6k	528.2k	171.3k
Average Query Length (words)	12.4	9.4	10.8	3.3	15.3	6.6
Average Document Length (words)	213.6	176.2	132.2	232.3	466.4	160.8
Relevant Document / Query	1.1	4.9	2.6	38.2	69.9	15.17

Table 5: Detailed statistics of the six subtasks included in the BEIR Benchmark as employed in our experiments. This table presents the number of queries, number of documents, average length of query and document, and the number of relevant documents for a query, for each subtask. Our experiments adopt the same preprocessing procedure as described in the GPL framework by Wang et al. 2021.

Retriever	Method	SF	SD	FQ	NF	RB	TC	Avg.
coCondenser	GPL	**64.7**	30.6	41.1	52.4	69.5	88.6	57.82
coCondenser	Ours: GPL+DADA	64.5	29.6	**42.2**	**52.5**	**70.0**	**91.0**	**58.3**
COCO-DR	GPL	66.3	**31.1**	**42.2**	52.3	**68.7**	93.1	**58.95**
COCO-DR	Ours: GPL+DADA	**66.5**	**30.5**	41.2	**53.8**	67.4	**93.4**	58.8
TAS-B	GPL	**63.6**	28.7	**40.8**	52.8	**68.0**	88.6	57.08
TAS-B	Ours: GPL+DADA	**63.6**	**30.0**	39.7	**54.5**	66.7	**89.2**	**57.28**
coCondenser	GPL+	59.5	28.7	**40.2**	**52.7**	65.3	**90.5**	56.15
coCondenser	Ours: GPL$^+$ + DADA	**60.1**	**30.4**	39.0	52.4	**67.2**	90.1	**56.53**

Table 6: Evaluation of **MRR@10** scores across different datasets. The best performance on each dataset for each retriever is highlighted in **bold**.

back. As we stated on §3.2.1, the observation feedback can be replaced by other statistics, we also experimented with *naver/splade-cocondenser-ensembledistil* and the result can be found on Table 9.

Retriever	Method	SF	SD	FQ	NF	RB	TC	Avg.
coCondenser	GPL	**65.2**	31.6	42.1	**52.9**	70.0	88.6	58.4
coCondenser	Ours: GPL+DADA	65.0	30.8	**43.2**	52.8	**70.5**	**91.0**	**58.88**
COCO-DR	GPL	66.8	**32.1**	**43.1**	53.0	**69.0**	93.1	**59.52**
COCO-DR	Ours: GPL+DADA	**67.0**	31.6	42.1	**54.3**	68.0	**93.5**	59.42
TAS-B	GPL	64.0	29.8	**41.6**	53.4	**68.5**	88.6	57.65
TAS-B	Ours: GPL+DADA	**64.1**	**31.2**	40.6	**55.0**	67.1	**89.2**	**57.87**
coCondenser	GPL+	60.0	29.8	**41.1**	**53.1**	65.9	**90.5**	56.73
coCondenser	Ours: GPL$^+$ + DADA	**60.6**	**31.4**	40.0	53.0	**67.6**	90.1	**57.12**

Table 7: Evaluation of **MRR@100** scores across different datasets. The best performance on each dataset for each retriever is highlighted in **bold**.

Retriever	Method	SF	SD	FQ	NF	RB	TC	Avg.
coCondenser	GPL	**70.4**	**23.6**	39.6	**30.6**	35.4	52.1	41.95
coCondenser	Ours: GPL+DADA	70.2	23.1	**40.7**	30.5	**35.5**	**53.4**	**42.23**
COCO-DR	GPL	72.3	**23.8**	**40.4**	31.2	35.0	**54.6**	**42.88**
COCO-DR	Ours: GPL+DADA	**72.5**	**23.8**	39.9	**31.4**	**35.3**	53.7	42.77
TAS-B	GPL	**69.4**	23.0	**39.2**	30.1	**33.6**	**50.9**	**41.03**
TAS-B	Ours: GPL+DADA	69.3	**23.5**	38.6	**30.4**	33.5	50.2	40.92
coCondenser	GPL+	66.2	22.7	**38.7**	**29.8**	**32.3**	50.6	40.05
coCondenser	Ours: GPL$^+$ + DADA	**66.7**	**23.7**	38.4	28.8	32.1	**52.0**	**40.28**

Table 8: Evaluation of **nDCG@100** scores across different datasets. The best performance on each dataset for each retriever is highlighted in **bold**.

Retriever	Method	SF	SD	FQ	NF	RB	TC	Avg.
coCondenser (Gao and Callan, 2022)	GPL	67.5	**17.1**	32.8	33.7	**42.7**	71.2	44.2
	Ours: GPL + DADA	**67.8**	16.9	**33.1**	**34.5**	41.7	**73.6**	**44.6**

Table 9: Comparative evaluation of nDCG@10 scores across different datasets. The experiment is conducted with ***naver/splade-cocondenser-ensembledistil*** as observation feedback.