HIL: Hybrid Isotropy Learning for Zero-shot Performance in Dense Retrieval

Jaeyoung Kim[1], **Dohyeon Lee**[2], **Seung-won Hwang**[12*]
[1]IPAI, [2]Computer Science and Engineering
Seoul National University
{jae.young, waylight3, seungwonh}@snu.ac.kr

Abstract

Advancements in dense retrieval models have brought ColBERT to prominence in Information Retrieval (IR) with its advanced interaction techniques. However, ColBERT is reported to frequently underperform in zero-shot scenarios, where traditional techniques such as BM25 still exceed it. Addressing this, we propose to balance representation isotropy and anisotropy for zero-shot model performance, based on our observations that isotropy can enhance cosine similarity computations and anisotropy may aid in generalizing to unseen data. Striking a balance between these isotropic and anisotropic qualities stands as a critical objective to refine model efficacy. Based on this, we present ours, a Hybrid Isotropy Learning (HIL) architecture that integrates isotropic and anisotropic representations. Our experiments with the BEIR benchmark show that our model significantly outperforms the baseline ColBERT model, highlighting the importance of harmonized isotropy in improving zero-shot retrieval performance.

1 Introduction

Recent advancements in information retrieval (IR) have seen a shift from classical techniques like BM25 to more advanced models that rely on dense vector representations. ColBERT (Khattab and Zaharia, 2020) is one such model utilizing multi-vector representations to capture token-level interactions, for often exceeding performance of single-vector dense models like DPR (Karpukhin et al., 2020). Despite its strengths, ColBERT lags behind BM25 and other methods in zero-shot tasks (Thakur et al., 2021), revealing a weakness in its representational strategy.

To improve representation quality from an isotropy perspective, methods can be categorized into approaches that promote isotropy and those that induce anisotropy. Specifically, approaches such as CosReg (Gao et al., 2019) promote isotropy, while I-STAR (Rudman and Eickhoff, 2023) induces anisotropy. Advocates for isotropy suggest it benefits cosine-based scoring (Jung et al., 2023; Li et al., 2020; Su et al., 2021), yet there are contrasting arguments that excessive isotropy might compromise a model's generalization abilities for unseen data (Zhu et al., 2018; Rudman and Eickhoff, 2023). However, these approaches have two limitations. **First**, the metrics used in their study are not suitable for IR as they calculate isotropy of query and passage as a single distribution, thus not taking into account the differences between the two distributions and their interactions. **Second**, these methods aim to make representations entirely isotropic or anisotropic, lacking the ability to achieve their balance.

Method	Interactive	Objective	
		Isotropic	Anisotropic
CosReg	✗	✓	✗
I-STAR	✗	✗	✓
Ours: ColBERT-HIL	✓	✓	✓

Table 1: Properties of methods with isotropy perspective. The Interactive column denotes whether the method considers the interaction between two distinct distribution. The Isotropic and Anisotropic columns indicate that the method make the representation isotropic and anisotropic, respectively.

To address these issues, we identify two challenges: (1) generalizing isotropy to IR, and (2) achieving an effective balance between isotropic and anisotropic properties in the representations. Table 1 illustrates the differences between our proposed method and baselines in the context of the two challenges. To address the first challenge, we introduce a new isotropy metric, Interactive Isotropy (InterIso), which reflects the interaction between two distributions. Additionally, we propose isotropic and anisotropic ColBERT models, namely ColBERT-iso and ColBERT-aniso, which leverage InterIso for regularization. To

*Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7889–7900

tackle the second challenge, we introduce Hybrid Isotropy Learning (HIL) architecture to integrate both isotropic and anisotropic modules.

For the first challenge, we propose, InterIso, measuring the interaction between two distributions. In IR, queries are typically shorter, and passages are more detailed and descriptive. However, traditional isotropy tools, such as Avg-Cos (Ethayarajh, 2019), IsoScore (Rudman et al., 2022), and Partition isotropy score (Mu et al., 2017), measure isotropy assuming that both query and passage representations originate from the same distribution. In contrast, InterIso computes the cosine similarity specifically between query and passage embeddings to capture query-passage interactions to reflect distributional differences.

For the second challenge, HIL proposes a hybrid model structure that combines the advantages of isotropy and anisotropy, consisting of two components: an isotropic lower layer and an anisotropic upper layer. We hypothesizes that maximizing the isotropy difference (ΔInterIso) between the isotropic and anisotropic components is our key contribution to harnessing their respective benefits. Specifically, we explore two strategies: an ensemble of separately trained models combined at score level and HIL approach that integrates both isotropic and anisotropic aspects within a single model. We find that promoting a larger ΔInterIso enhances zero-shot performance, with HIL proving superior as it encourages an isotropy-anisotropy balance throughout the model layers.

To demonstrate the effectiveness of our approach for Dense Retrieval (DR) in zero-shot setting, we experiment with BEIR benchmarks. As shown in Table 2, ColBERT-HIL significantly outperforms BM25 and ColBERT by +2.22%, +2.7% in full-ranking retrieval, respectively. It represents the importance of hybrid isotropy for better zero-shot performance. Our contributions can be summarized as the following.

- Identifying challenges in adapting isotropy insights to IR models.
- Developing InterIso, a new metric to assess isotropy in IR more accurately.
- Introducing ColBERT-HIL to achieve a synergistic mix of isotropic and anisotropic representations.
- Demonstrating the benefits of ColBERT-HIL for zero-shot performance via BEIR benchmark results.

2 Related Work

2.1 Dense retrieval

The BERT (Devlin et al., 2019) models have shown promise for dense retrieval tasks, where both queries and documents are represented through embeddings. Dense retrieval models fall into two categories: single-vector and multi-vector models. Our focus is on the latter, where individual embeddings represent every term within queries and documents. In this scenario, query token embeddings are used to retrieve the nearest document token embeddings through approximate nearest neighbor search (ANN). In the case of ColBERT, the documents retrieved through ANN are reranked using late interaction as defined below, which is also applied during training.

$$\text{score}_{\text{late}}(q,d) = \sum_{i \in [|E_q|]} \max_{j \in [|E_d|]} E_{q_i} \cdot E_{d_j}^T \quad (1)$$

where q and d represent a query and a document respectively, and $E_{(.)}$ represents token embedding matrix of query or document.

2.2 Zero-shot retrieval

Zero-shot retrieval is a paradigm in IR that focuses on the ability to retrieve relevant information for queries which were not encountered during the training phase. The late-interaction model, ColBERT performs a bit weaker than BM25 and the other models on BEIR benchmark (Thakur et al., 2021), which implies a weakness in its representations. While Jung et al. (2023) study isotropic representations using post-processing with out-of-distribution datasets in DR, they do not conduct it in full-ranking and focus solely on isotropic representations, rather than hybrid isotropy. Given our emphasis on isotropy, there is no need to be concerned about other late-interaction models that do not consider isotropy.

2.3 Isotropy

In NLP, Isotropy refers to the measure of how evenly distributed the contextualized representations are. We briefly review the commonly used metrics and methods in Semantic Textual Similarity (STS).

To measure isotropy, **Avg-Cos** (Ethayarajh, 2019) computes the average cosine similarity score

among representations. **IsoScore** (Rudman et al., 2022) measures the distance between the covariance matrix of the data and the identity matrix. **Partition** isotropy score, as defined by Mu et al. (2017), involves a specific quotient related to the partition function initially proposed by Arora et al. (2016).

To control isotropy, Rudman and Eickhoff (2023); Gao et al. (2019) suggest regularization term, I-STAR and CosReg, to control isotropy during training. While I-STAR is designed to be anisotropic and CosReg is intended to be isotropic in its design, both can induce either isotropy or anisotropy by adjusting λ.

3 Proposed Method

We introduce InterIso measuring isotropy in Section 3.1, and then discuss ensemble model using InterIso in Section 3.2. We then propose ColBERT-HIL mitigating the limitations of ensemble in Section 3.3.

3.1 Interactive Isotropy (InterIso)

Traditional isotropy metrics might distort the evaluation due to assuming that both query and passage representations come from the same distribution. Thus, to extend isotropy to IR, it is essential to consider the interaction between query and passage distributions. In this regard, we introduce Interactive Isotropy (InterIso) to assess isotropy for query and passage pairs. It can be formulated as follows:

$$\text{InterIso}(q,d) = \frac{1}{|E_q||E_d|} \sum_{i\in[|E_q|]} \sum_{j\in[|E_d|]} E_{q_i} \cdot E_{d_j}^T \tag{2}$$

where q and d represent a query and a passage respectively, and $E_{(\cdot)}$ represents token embedding matrix of query or passage.

InterIso is a simple yet effective measure of isotropy between query and passage distributions. InterIso, when closer to 0, signifies isotropic space, where passage token embeddings make a sparse space for each query token embedding. Conversely, a value closer to 1 means that passage token embeddings are dense for each query token embedding, indicating anisotropic space. Note that Eq. (2) computes similarity only between query and passage tokens, excluding comparisons within the individual query or passage, to capture the interaction between query-passage distribution.

3.2 Ensemble for Hybrid Isotropy

To leverage the advantages of both isotropy and anisotropy in ensemble model, it may be crucial to increase ΔInterIso, which denotes the difference in InterIso between isotropic and anisotropic modules. For this purpose, InterIso metric can be directly employed as a loss term for each module. Given triples $\langle q, d^+, d^- \rangle$, we can formulate late-interaction loss, $\mathcal{L}_{\text{late}}$ term as follows:

$$\mathcal{L}_{\text{late}} = \text{CE}(\text{score}_{\text{late}}(q,d^+),\ \text{score}_{\text{late}}(q,d^-))$$

where CE represents cross-entropy function, d^+ and d^- represent a positive and negative passage for given query q. The loss, $\mathcal{L}_{\text{InterIso}}$ can be formulated as follows:

$$\mathcal{L}_{\text{reg}}(q,d) = \begin{cases} \text{InterIso}(q,d) & \lambda < 0 \\ \text{abs}(\text{InterIso}(q,d)) & \lambda > 0 \end{cases}$$

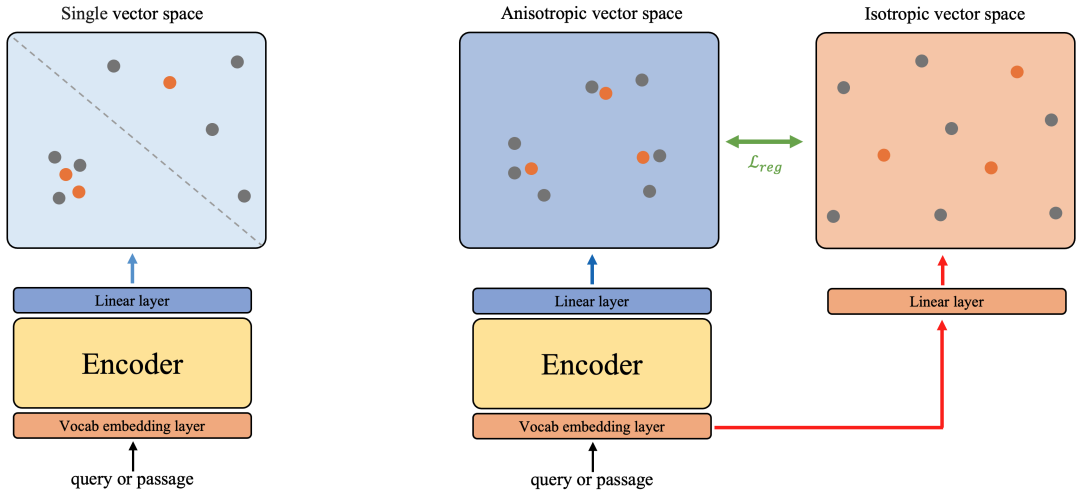
$$\mathcal{L}_{\text{InterIso}} = \mathcal{L}_{\text{late}} + \lambda\, \mathcal{L}_{\text{reg}}$$

where abs$(\cdot)$ represents absolute value. As InterIso closer to 0 represents strong isotropy, we take the absolute value of InterIso for $\lambda > 0$. The isotropy of representations is controlled by the sign of λ, where negative values make it more anisotropic and positive values make it more isotropic. We choose ColBERT-iso and ColBERT-aniso as isotropic and anisotropic modules, respectively, based on their nDCG@10 performance for $\lambda > 0$ and $\lambda < 0$ on MSMARCO dev dataset. ColBERT-λ is a simple $\lambda-$ensemble combining these modules as a baseline, and detailed implementations of ColBERT-cosreg and ColBERT-istar using CosReg and I-STAR regularization are described in Section A.3.

3.3 Hybrid Isotropy Learning (HIL)

While ColBERT-λ can enhance ΔInterIso by adjusting λ of each modules, extreme changes in λ might disrupt the vector spaces learned in the pretraining step. To tackle this issue, we propose ColBERT-HIL with Hybrid Isotropy Learning (HIL) architecture, achieving both isotropic and anisotropic representation within a single model.

Inspired by the findings of Ethayarajh (2019), which demonstrate that BERT tends to learn isotropic embeddings in its lower layers and anisotropic embeddings in its upper layers, we adopt a similar strategy. As shown in Figure 1, ColBERT employs a single vector space, whereas ColBERT-HIL utilizes two, separately learning



(a) ColBERT (b) ColBERT-HIL

Figure 1: Overview of ColBERT (left) and ColBERT-HIL (right). (b) shows its anisotropic and isotropic vector spaces represented by blue box and orange box, respectively. In vector spaces, orange and gray dots indicate query and passage token embeddings, respectively.

isotropic and anisotropic representations to harness their respective benefits. Specifically, we take isotropic representations (L) which are sequentially passed through the vocab embedding layer and a linear layer aimed at reducing the embedding dimension. Furthermore, anisotropic representations (H) are obtained by projecting the last hidden state, which aligns with the ColBERT token representations.

To formulate the loss function for learning each representation, we describe further details. Let s denote either a query or a passage.

$$\text{Input}(s) = \text{[CLS] [S] Tokenizer}(s) \text{ [SEP]}$$

where [S] becomes [Q] for a query and [D] for a passage, depending on s. Then, L_s and H_s can be formulated as follows:

$$L_s = \text{Linear}(\text{Vocab}(\text{Input}(s))) \in \mathbb{R}^{|s| \times 128}$$
$$H_s = \text{Linear}(\text{BERT}(\text{Input}(s))) \in \mathbb{R}^{|s| \times 128}$$

where Vocab represents vocab embedding layer. As L_s represents a projected Vocab embedding, it may contain potential noise from less important terms. To address this, we multiply the inverse document frequency (IDF) of the query term, approximating term importance (Formal et al., 2021), by the maximum value in Eq. (1). Let $M(q_i, d)$ denote the maximum similarity between query term q_i and passage token embeddings. Then, the scores of L and H for q_i can be formulated as follows:

$$M(q_i, d) = \max_{j \in [|E_d|]} E_{q_i} \cdot E_{d_j}^T$$
$$S_L(q_i, d) = \text{IDF}(q_i) \cdot M(q_i, d)$$
$$S_H(q_i, d) = M(q_i, d)$$

However, this multiplication introduces a difference in scale between $S_L(q_i, d)$ and $S_H(q_i, d)$. To tackle this scaling issue, we compute mean and standard deviation of them using samples in the same batch, and apply z-score normalization. Then, late-interaction score and loss for L and H can be formulated as follows:

$$S'_{(\cdot)}(q_i, d) = \frac{S_{(\cdot)}(q_i, d) - m_{(\cdot)}}{\sigma_{(\cdot)}}$$
$$\text{score}_{\text{late}_{(\cdot)}}(q, d) = \sum_{i \in [|E_q|]} S'_{(\cdot)}(q_i, d)$$
$$\mathcal{L}_{\text{late}_{(\cdot)}} = \text{CE}(\text{score}_{\text{late}_{(\cdot)}}(q, d^+),\ \text{score}_{\text{late}_{(\cdot)}}(q, d^-))$$

where $m_{(\cdot)}$ and $\sigma_{(\cdot)}$ represent mean and standard deviation, respectively. To combine two late-interaction losses, we simply add them together.

$$\mathcal{L}_{\text{HIL}} = \mathcal{L}_{\text{late}_L} + \mathcal{L}_{\text{late}_H}$$

Although L_s and H_s already possess isotropy and anisotropy as suggested by Ethayarajh (2019), we expect that incorporating InterIso as a regularization term can further enhance their isotropic difference. Therefore, we revise $\mathcal{L}_{\text{HIL}}$ by introducing

Model(→)	Baselines		Hybrid Isotropy Models			
Dataset(↓)	BM25	ColBERT	ColBERT-cosreg	ColBERT-istar	ColBERT-λ	ColBERT-HIL †(ours)
MS MARCO (dev)	22.84	41.32	41.38	41.89	**42.17**	42.04
TREC-COVID	59.35	66.65	40.43	65.06	65.57	**73.21**
BioASQ	**52.25**	43.58	44.32	43.74	43.54	45.25
NFCorpus	**32.06**	28.82	28.26	28.89	29.15	31.33
NQ	30.55	51.99	42.5	51.79	**52.28**	51.08
HotpotQA	63.29	60.09	58.74	60.14	59.82	**67.11**
FiQA-2018	23.61	29.83	28.57	29.11	29.3	**31.39**
Signal-1M (RT)	**33.04**	27.61	25.09	28.56	27.44	30.09
TREC-NEWS	**39.52**	35.79	32.44	36.6	37.43	37.35
Robust04	40.7	36.8	36.61	37.2	38.41	**42.18**
ArguAna	**39.7**	29.8	29.47	30.17	29.54	33.04
Touché-2020	**44.25**	21.62	13.89	20.96	21.97	25.53
CQADupStack	30.21	34.69	33.74	34.88	34.9	**36.19**
Quora	78.84	84.76	82.84	84.72	**85.39**	85.07
DBPedia	31.78	38.01	36.27	39.74	39.12	**40.04**
SCIDOCS	14.9	15.15	13.92	15.13	14.81	**15.8**
FEVER	65.13	72.39	72.93	74.14	73.9	**76.35**
Climate-FEVER	16.51	15.08	14.78	15.77	17.02	**17.33**
SciFact	**67.89**	62.29	62.54	62.18	62.63	65.27
Avg. Performance vs. BM25	-	-0.48%	-3.68%	-0.27%	-0.08%	**+2.22%**
Avg. Performance vs. ColBERT	+0.48%	-	-3.2%	+0.21%	+0.4%	**+2.7%**

Table 2: Full-ranking nDCG@10 performances of BM25, ColBERT, and hybrid isotropy models on MS MARCO and BEIR benchmarks. The best performing results are highlighted in bold, and second best performing results are underlined. The Avg. performance does not include MSMARCO, and † represents the results with the p-value $<$ 0.01 in comparison to ColBERT.

regularization terms as follows:

$$\mathcal{L}_{\text{reg}_L} = -\text{abs}(\text{InterIso}_L(q,d))$$
$$\mathcal{L}_{\text{reg}_H} = \text{InterIso}_H(q,d)$$
$$\mathcal{L}_{\text{HIL}} = \mathcal{L}_{\text{late}_L} + \mathcal{L}_{\text{late}_H} + \lambda\,(\mathcal{L}_{\text{reg}_L} + \mathcal{L}_{\text{reg}_H})$$

where $\text{InterIso}_L(q,d)$ and $\text{InterIso}_H(q,d)$ represent InterIso for L_s and H_s, respectively. In $\mathcal{L}_{\text{reg}_L}$ term, we take the absolute value to make $\text{InterIso}_L(q,d)$ close to 0, and use a minus sign to increase the difference from $\text{InterIso}_H(q,d)$. As our objective is to enhance ΔInterIso, we conducted experiments with a negative value of λ. At inference step, we compute $\text{score}_{\text{late}_L}(q,d)$ and $\text{score}_{\text{late}_H}(q,d)$, and then simply add them together.

4 Experiment

4.1 Experimental Setting

Dataset and Evaluation Metric To validate our approach, we use two datasets, MS MARCO-passage (Nguyen et al., 2016) for training and BEIR (Thakur et al., 2021) for evaluating zero-shot performance. Our primary focus lies in improving the Normalized Discounted Cumulative Gain (nDCG) metric for full-ranking retrieval on BEIR benchmark.

Implementation We follow the training settings described in (Khattab and Zaharia, 2020) as we implement our approach using ColBERTv1 as the backbone. All ColBERT model variations in this paper are trained under the same settings. For BM25, we use open-source implementation provided by Pyserini[1]. More detailed hyperparameters are described in Section A.3.

4.2 Experimental Results

Research Questions To evaluate the effectiveness of our method, we address the following research questions:

- RQ1: Does increasing ΔInterIso enhance zero-shot performance and why?
- RQ2: Does HIL architecture mitigate a limitation of $\lambda-$ensemble?
- RQ3: Does each regularization term in HIL improve ΔInterIso?
- RQ4: Is HIL architecture generalizable?

[1] https://github.com/castorini/pyserini

Model	BEIR	ΔInterIso
ColBERT-cosreg	45.54	-0.01
ColBERT-istar	45.75	0.01
ColBERT-λ	46.12	0.5
ColBERT-HIL	**46.7**	**0.99**

Table 3: nDCG@10 performance in re-ranking top100 passages from BM25 and ΔInterIso for various isotropic and anisotropic modules.

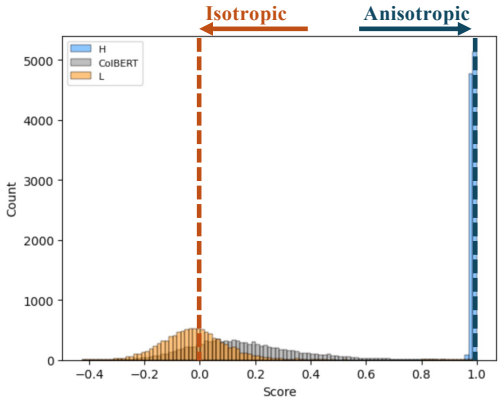


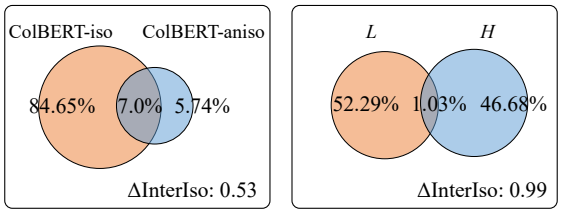
Figure 2: Score distributions for ColBERT, L and H in ColBERT-HIL. Dashed lines represent ideal isotropy (orange) and anisotropy (blue).

4.2.1 RQ1: ΔInterIso for zero-shot performance

In this section, we demonstrate that augmenting ΔInterIso yields improvements in zero-shot performance, while also conducting an analysis of whether isotropic and anisotropic modules are performing as intended.

Zero-shot performance We first evaluate zero-shot performance in full-ranking, as shown in Table 2. Remarkably, our model (ColBERT-HIL) exhibits the most superior performance in this context. In delving deeper into the relationship between ΔInterIso and zero-shot performance, we analyze re-ranking performance using various ensemble models. Table 3 indicates that ColBERT-λ and ColBERT-HIL, utilizing InterIso as loss term, achieve the second-best and best performance, respectively. This suggests enhanced zero-shot performance with an increase in ΔInterIso.

Isotropy and Anisotropy are Well-Represented To affirm the successful representation of isotropic and anisotropic spaces in ColBERT-HIL, by L and H modules, respectively, we compute the cosine similarity scores between query and passage token embeddings. As depicted in Figure 2, the score distributions of L (orange curve) and H (blue curve) noticeably shift towards 0 and 1, respectively, when compared to the baseline ColBERT (grey curve). This empirical evidence supports the assertion that the isotropic and anisotropic modules within ColBERT-HIL effectively learn their respective spaces for IR.



(a) ColBERT-λ (b) ColBERT-HIL

Figure 3: Venn diagram of retrieved passages in first-stage. The orange and blue indicate passage sets retrieved by isotropic and anisotropic modules, respectively.

Isotropy and Anisotropy are Complementary To explore the complementarity between isotropic and anisotropic modules complementarity inherent in isotropic and anisotropic modules, we present a Venn diagram depicting the retrieved passages for each query in the first stage, as depicted in Figure 3.

In both ColBERT-λ and our proposed model (ColBERT-HIL), we observe that isotropic and anisotropic modules discover mostly distinct passages but with some duplications (shown as intersection in the diagram): while ColBERT-λ retrieves 7.0% duplicated passages with a ΔInterIso of 0.53, ColBERT-HIL retrieves only 1.03% duplicated passages, demonstrating a significant reduction with a higher ΔInterIso of 0.99. This outcome presents the pronounced complementary advantage achieved by increasing ΔInterIso.

Meanwhile, in the case of ColBERT-λ, the majority of passages (91.65%) are retrieved by ColBERT-iso. This limitation arises because ColBERT-aniso may retrieve passage tokens from the same passage for each query token, owing to its anisotropic distribution. In contrast, ColBERT-HIL demonstrates a more balanced retrieval, with comparable numbers of passages retrieved using its isotropic and anisotropic modules. This balanced approach mitigates the bias observed in ColBERT-λ's retrieval mechanism, where a predominant reliance on ColBERT-iso occurs.

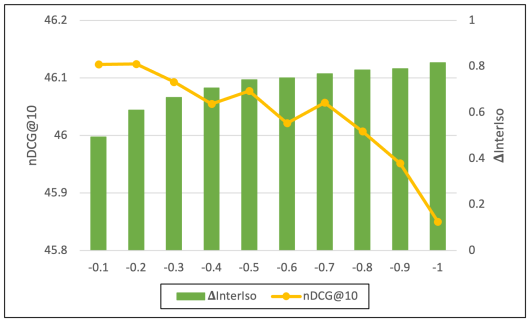


Figure 4: Performance drop when decreasing λ for ColBERT-aniso. Yellow line and green bar indicate nDCG@10 in re-ranking on BEIR and ΔInterIso of ColBERT-λ, respectively. The x-axis represents the λ for regularization term of ColBERT-aniso.

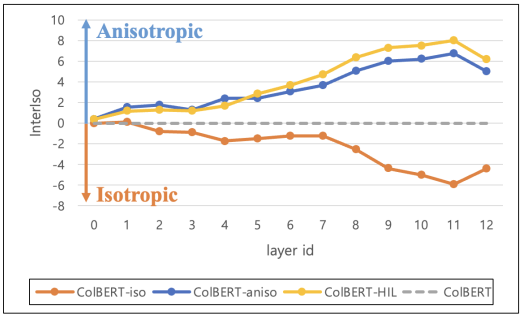


Figure 5: Differences in the log-scale of InterIso compared to ColBERT for each layer.

Model	MSMARCO	BEIR
Baselines		
BM25	59.24	87.1
ColBERT	66.15	82.15
Anisotropic module		
ColBERT-aniso	64.42	76.54
H	**65.01**	**81.22**
Hybrid model		
ColBERT-λ	66.51	82.4
ColBERT-HIL	**67.37**	**83.85**

Table 4: Re-ranking nDCG@10 performance on BM25-Easy dataset. The H represents anisotropic module of ColBERT-HIL.

Model	MSMARCO	BEIR	ΔInterIso
ColBERT-HIL w/o reg_L, reg_H	41.76	43.6	0.24
ColBERT-HIL w/o reg_H	41.96 (+0.2)	43.84 (+0.24)	0.29 (+0.05)
ColBERT-HIL w/o reg_L	42.2 (+0.44)	44.54 (+0.94)	0.94 (+0.7)
ColBERT-HIL	**42.34** (+0.58)	**44.8** (+1.2)	**0.99** (+0.75)

Table 5: Ablation study for regularization terms. The number indicates full-ranking nDCG@10 performance for each dataset.

4.2.2 RQ2: Mitigating the limitation of ColBERT-λ

In this section, we reveal a limitation of ColBERT-λ, and HIL architecture mitigates this issue.

To optimize the complementarity between isotropy and anisotropy, we observe zero-shot performance and ΔInterIso of ColBERT-λ across various λ values for each module. As depicted in Figure 4, a decline in performance is evident when reducing the λ of ColBERT-aniso to enhance its anisotropy while maintaining ColBERT-iso fixed. This suggests a limitation in $\lambda-$based tuning for finding an optimal ensemble.

We also examine isotropy and anisotropy in each layer, to confirm whether it accords with finding from Ethayarajh (2019): BERT is known to acquire isotropic embeddings in its lower layers and anisotropic embeddings in its upper layers. Our finding illustrated in Figure 5 is consistent: Lower layers, isotropic and containing lexical information, are influenced by the regularization term, leading to induced anisotropy in upper layers. As a result, InterIso, calculated for each layer in the figure, is significantly higher values in the lower layers of ColBERT-aniso, indicating a potential loss of lexical information.

For a more detailed analysis, we introduce the BM25-Easy dataset based on the top 50% of BM25 nDCG@10 scores for all queries. The evaluation of performances on BM25-Easy in Table 4 reveals that ColBERT-aniso experiences a performance decrease of -5.61% compared to ColBERT, signifying a disturbance in lexical information.

On the other hand, the HIL architecture circumvents these issues. The objectives of the $\mathcal{L}_{\text{reg}L}$ and $\mathcal{L}_{\text{reg}H}$ terms in the ColBERT-HIL loss function directly aim to enhance ΔInterIso during training, inducing isotropy and anisotropy in its lower and upper layers, respectively. In Figure 5, ColBERT-HIL exhibits more isotropy than ColBERT-aniso up to the fifth layer and becomes increasingly anisotropic from the sixth layer onward. Furthermore, in Table 4, the anisotropic module H outperforms ColBERT-aniso by +4.68%, and ColBERT-HIL surpasses ColBERT-λ by +1.45% on BM25-Easy. This suggests that ColBERT-HIL effectively mitigates the disturbance of lexical information.

Model	MSMARCO	BEIR
ColBERT	41.66	42.32
ColBERT-λ	42.03 (+0.37)	42.52 (+0.2)
ColBERT-HIL	**42.34** (+0.68)	**44.8** (+2.48)

Table 6: Re-ranking nDCG@10 performance for same passages, retrieved in first-stage by ColBERT-HIL.

Model	MSMARCO	BEIR
CITADEL	35.86	38.03
CITADEL-HIL	**36.84** (+0.98)	**40.32** (+2.29)

Table 7: Re-ranking nDCG@10 performance for each dataset. CITADEL-HIL represents CITADEL trained with HIL architecture.

4.2.3 RQ3: Ablation Study

We conduct an ablation study on both the MSMARCO and BEIR datasets, as illustrated in Table 5. Each regularization term is individually incorporated into ColBERT-HIL without any other regularization terms.

Our findings reveal that both $\mathcal{L}_{\text{reg}L}$ and $\mathcal{L}_{\text{reg}H}$ contribute to enhancements in ΔInterIso and overall performance across both datasets. Notably, $\mathcal{L}_{\text{reg}L}$, designed to induce isotropy in the lower layers, leads to an improvement in ΔInterIso. This holds true even when the lower layers are already strong isotorpic when trained without any regularization terms.

It is worth noting that the individual contributions of both $\mathcal{L}_{\text{reg}L}$ and $\mathcal{L}_{\text{reg}H}$ are evident in the observed improvements, underlining the effectiveness of each term in refining the performance of ColBERT-HIL on both datasets.

4.2.4 RQ4: Generalizability of HIL

This section shows how our proposed approach, the HIL structure utilizing the InterIso metric, generalizes to re-ranking task, to a more recent late-interaction model.

Re-ranking task To assess the re-ranking capabilities of ColBERT, ColBERT-λ, and ColBERT-HIL, we conduct re-ranking on the same passages retrieved in the first stage by ColBERT-HIL, as presented in Table 6. Despite employing an equal number of passages for re-ranking, both ColBERT-HIL and ColBERT-λ surpass ColBERT in performance. This suggests the advantages of leveraging both isotropic and anisotropic representations for re-ranking. Additionally, ColBERT-HIL significantly enhances zero-shot performance by approximately 12 times compared to ColBERT-λ, underscoring the effectiveness of the HIL framework for zero-shot retrieval.

Recent late-interaction model To validate the generalizability of HIL for more recent late-interaction model, other than ColBERT, we apply HIL to the recent late-interaction model, CITADEL (Li et al., 2023). As shown in Table 7, CITADEL-HIL, trained with the HIL architecture, outperforms CITADEL by +0.98% and +2.29% on MSMARCO and BEIR, respectively. The improvement in BEIR performance is twice that observed in MSMARCO, aligning with the consistent results observed with ColBERT and highlighting the versatility of HIL across different retrieval models.

5 Conclusion

Addressing the challenge of applying prior insights on isotropy to IR, we introduce InterIso, a metric designed to quantify isotropy between queries and passages in DR settings. Leveraging InterIso, we propose HIL architecture that seamlessly integrates isotropic and anisotropic representations. Our experiments with BEIR benchmark demonstrate that ColBERT-HIL significantly outperforms the baselines, ColBERT and BM25. It indicates the pivotal role of harmonized isotropy in enhancing zero-shot retrieval performance.

6 Limitations

To the best of our knowledge, we are the first to achieve hybrid isotropy by combining isotropic and anisotropic embeddings at the score-level. While this approach is straightforward and easy to apply, adopting more sophisticated methods to combine isotropy and anisotropy could potentially enhance performance beyond simple score-level fusion. We remain this task for future research endeavors.

Acknowledgements

This research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was also partially supported by

IITP grant funded by MSIT [No.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data and No.2021-0-01343-004, Artificial Intelligence Graduate School Program (Seoul National University)].

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A white box analysis of colbert. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 257–263. Springer.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Euna Jung, Jungwon Park, Jaekeol Choi, Sungyoon Kim, and Wonjong Rhee. 2023. Isotropic representation can improve dense retrieval. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 125–137. Springer.

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.

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.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. Citadel: Conditional token interaction via dynamic lexical routing for efficient and effective multi-vector retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11891–11907.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

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

William Rudman and Carsten Eickhoff. 2023. Stable anisotropic regularization. In *The Twelfth International Conference on Learning Representations*.

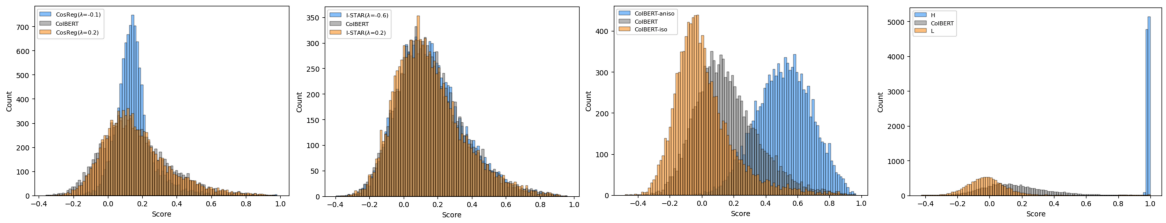
William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

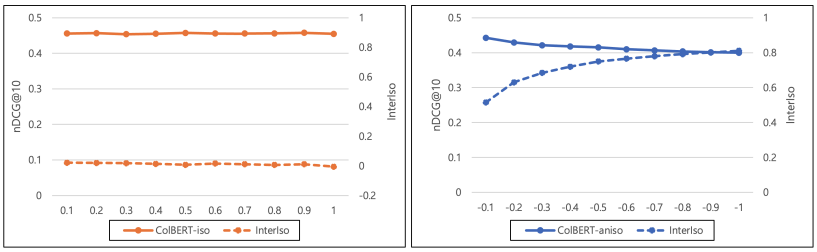
Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 873–876.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. 2018. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*.



(a) ColBERT-cosreg (b) ColBERT-istar (c) ColBERT-sensim (d) ColBERT-HIL

Figure 6: Score distributions for hybrid isotropy models. Orange and Blue represent isotropic and anisotropic modules, respectively, and Gray indicates ColBERT.



(a) ColBERT-iso (b) ColBERT-aniso

Figure 7: Re-ranking nDCG@10 performance on BEIR and InterIso for ColBERT-iso (left) and ColBERT-aniso (right). Solid and dashed lines represent nDCG@10 and InterIso, respectively. The x-axis indicates λ for each module.

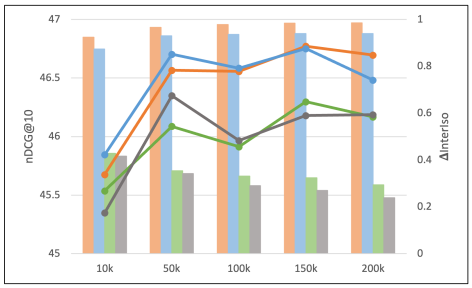


Figure 8: Re-ranking nDCG@10 performance of ColBERT-HIL on BEIR and ΔInterIso during training. The x-axis indicates training steps, and Red, Blue, Green, and Gray represent ColBERT-HIL, ColBERT-HIL w/o reg_L, ColBERT-HIL w/o reg_H, and ColBERT-HIL w/o reg_L, reg_H, respectively.

A Appendices

A.1 Isotropy

Metrics **Avg-Cos** (Ethayarajh, 2019) computes the average cosine similarity score as follows:

$$\text{Avg-Cos}(s) = \frac{1}{|E_s|(|E_s|-1)} \sum_{i \neq j} E_{s_i} \cdot E_{s_j}^T$$

where s indicates representations. **IsoScore** (Rudman et al., 2022) measures the distance between the covariance matrix of the data and the identity matrix as follows:

$$\delta(X) = \frac{||\hat{\Sigma}_D - \mathbf{1}||}{\sqrt{2(n-\sqrt{n})}}$$

$$\phi(X) = \frac{(n - \delta(X)^2(n-\sqrt{n}))^2}{n^2}$$

$$\text{IsoScore}(X) = \frac{(n \cdot \phi(X) - 1)}{(n-1)}$$

where X, n, and $\hat{\Sigma}_D$ indicate representations, number of components of PCA for X, and normalized diagonal of the covariance matrix of first n principal components. **Partition isotropy score**, as defined by Mu et al. (2017), can be formulated as follows:

$$Z(C) = \sum_{x \in X} \exp(c^T x)$$

$$I(X) = \frac{\min_{c \in C} Z(c)}{\max_{c \in C} Z(c)}$$

where X indicates the representations and c is chosen from the eigenspectrum of XX^T. Avg-Cos closer to 0 indicates strong isotropy, while a value of 1 suggests anisotropy. For IsoScore and Partition isotropy score, 0 represents anisotropy, whereas 1 represents isotropy.

Methods Rudman and Eickhoff (2023); Gao et al. (2019) propose regularization term, I-STAR and CosReg, to adjust isotropy during training. Their loss function can be formulated as follows:

$$\mathcal{L}_{\text{I-STAR}} = \mathcal{L}_{\text{CE}} + \lambda \cdot (1 - \text{IsoScore}(\widetilde{X}, C_i))$$

$$\mathcal{L}_{\text{CosReg}} = \mathcal{L}_{\text{CE}} + \lambda \frac{1}{M^2} \sum_{i}^{M} \sum_{j \neq i} \hat{x}_i^T \hat{x}_j$$

where $\mathcal{L}_{\text{CE}}$ represents cross-entropy loss, $\widetilde{X} = \bigcup_{l=1}^{n} X_l$ denotes the union of all hidden states from a network with n layers and C_i is the shrinkage covariance matrix for epoch i of training. IsoScore measures isotropy of $\widetilde{X}$ and C_i. For CosReg, $\hat{x}_i = \frac{x_i}{||x||}$ and $\{x_1, x_2, ..., x_M\}$ denotes the mini-batch representation obtained from the last hidden layer.

A.2 Experiment Setting

MS MARCO-Passage [2] (Nguyen et al., 2016) This dataset offers a collection of 8.8 million passages with the labels sourced from the Bing search engine. Since the relevance labels for the official test set are not accessible to the public, we used only the training set to train the model and evaluated it on the development set.

BEIR [3] (Thakur et al., 2021) Benchmarking-IR (BEIR) serves as a comprehensive and diverse evaluation benchmark for zero-shot information retrieval. It encompasses 18 datasets spanning a range of text retrieval tasks and domains.

Evaluation Metric We evaluate the performance on the BEIR benchmark using the Normalized Discounted Cumulative Gain (nDCG) metric. To compute this metric, we employ the pytrec_eval (Van Gysel and de Rijke, 2018) Python library.

A.3 Implementation

Variations of ColBERT model We follow the training settings described in Khattab and Zaharia (2020), using a learning rate of 3×10^{-6} and a batch size of 32. The models are trained for 200k iterations on the MS MARCO-Passage dataset, with query and passage lengths set to 32 and 180. For ColBERT encoder, we utilize the base version (Uncased) of BERT (Devlin et al., 2019). For hyper-parameter, we search λ based on nDCG@10 using MS MARCO-Passage dev dataset in a range of [-1, 1] with a step size of 0.1. The best configuration λ was -0.1, 0.4 and -0.3 for ColBERT-aniso, ColBERT-iso and ColBERT-HIL, respectively.

Regularization term We revise CosReg and I-STAR, as regularization for DR. Given triples $\langle q, d^+, d^- \rangle$, we replace $\mathcal{L}_{\text{CE}}$ term, which is mentioned in Section A.1, with $\mathcal{L}_{\text{late}}$ as follows:

$$\mathcal{L}_{\text{late}} = \text{CE}(\text{score}_{\text{late}}(q, d^+),\ \text{score}_{\text{late}}(q, d^-))$$

where d^+ and d^- represent a positive and negative passage for given query q. Since both I-STAR and CosReg regularization terms have the same role, we explore two terms with hyper-parameter λ. The loss function can be formulated as follows:

$$\mathcal{L}_{\text{reg}}(s) = \begin{cases} 1 - \text{IsoScore}(X, I) \\ \frac{1}{M^2} \sum_{i}^{M} \sum_{j \neq i} X_i^T X_j \end{cases}$$

$$\mathcal{L}_{\{\text{CosReg} \mid \text{I-STAR}\}} = \mathcal{L}_{\text{late}} + \frac{\lambda}{2} \cdot (\mathcal{L}_{\text{reg}}(q) + \mathcal{L}_{\text{reg}}(d))$$

where s denotes either a query or a passage, and X and I represent projected last hidden state of s and identity matrix, respectively. The other notations are the same as described in Section A.1. We compute regularization term for queries and passages separately, recognizing their distinct distributions. The optimal configurations for λ were 0.2 and -0.1 for isotropic and anisotropic modules of CosReg, and 0.2 and -0.6 for isotropic and anisotropic modules of I-STAR.

CITADEL We use the open-source dpr-scale[4] to implement CITADEL (Li et al., 2023) and CITADEL-HIL. When implementing CITADEL-HIL, we L2 normalize token embedding and apply softmax operation to token-level router representations to prevent the divergence of InterIso values. We set λ, query and passage length to -0.3, 32 and 180, respectively, and train for 10 epochs.

A.4 Experiment details

Score distributions The score distributions for hybrid isotropy models are depicted in Figure 6. These scores are computed using query and passage pairs retrieved by BM25 from MSMARCO dev dataset. ColBERT-InterIso and ColBERT-HIL effectively learn isotropic and anisotropic spaces,

[2] https://github.com/microsoft/MSMARCO-Passage-Ranking

[3] https://github.com/beir-cellar/beir

[4] https://github.com/facebookresearch/dpr-scale

Model	BEIR
ColBERTv2	46.4
ColBERTv2-HIL	**46.61** (+0.21)

Table 8: nDCG@10 scores in full-ranking when using ColBERTv2 as a backbone model.

Model	BEIR
ColBERT	42.23
ColBERT-HIL	**45.25** (+3.02)

Table 9: nDCG@10 scores in full-ranking. ColBERT and ColBERT-HIL are trained with BEIR-hyperparameters.

whereas ColBERT-cosreg and ColBERT-istar struggle to learn distinct vector spaces. We can also see ColBERT-iso and ColBERT-aniso achieve the best isotropy and anisotropy, respectively, except for L and H.

Limitation of Ensemble To reveal limitations of ColBERT-InterIso, we examine zero-shot performance and InterIso of ColBERT-iso and ColBERT-aniso for various λ. As shown in Figure 7, we observe a performance drop in ColBERT-aniso when decreasing λ, while ColBERT-iso maintains consistent performance.

Ablation study for training steps We conduct an ablation study on training steps, as shown in Figure 8. It shows that L and H are close to each other when trained without any regularization term. While ΔInterIso for ColBERT-HIL w/o reg_H is larger than ColBERT-HIL w/o $\text{reg}_L, \text{reg}_H$ at the 200k step, it still decreases during training due to reg_L only inducing isotropy in its lower layers, which already exhibit strong isotropy. However, ΔInterIso for ColBERT-HIL consistently increases during training, signifying that reg_L and reg_H effectively enhance the isotropic difference between them.

A.5 Additional evaluations

ColBERTv2 as a backbone We adopted ColBERTv1 as a backbone model since ColBERTv2 is not suitable for separately observing the two types of scores in HIL: one from isotropic space and another from anisotropic space. However, in this section, we confirm the performance gain based on ColBERTv2. We used the public checkpoint[5] for ColBERTv2 and trained ColBERTv2-HIL, initializing it from a pretrained checkpoint which trained with one hard negative example over 150 steps. Despite ColBERTv2 being trained with 31 hard negative examples, we observe in Table 8 that ColBERTv2-HIL still outperforms it.

BEIR-hyperparameters We used the hyperparameters from the ColBERTv1 paper (Khattab and Zaharia, 2020) since we utilized ColBERTv1 as a backbone model. This slightly differs from the hyperparameters used in the BEIR paper (Thakur et al., 2021). In ColBERTv1 settings, the sequence length is 180, while the BEIR paper used a sequence length of 300. To verify the effect of hyperparameters, we trained ColBERT and ColBERT-HIL with BEIR-hyperparameters. Table 9 demonstrates that we can achieve a similar performance gain when using BEIR-hyperparameters.

[5] https://github.com/stanford-futuredata/ColBERT