USTAD: Unified Single-model Training Achieving Diverse Scores for Information Retrieval

Seungyeon Kim [*1] **Ankit Singh Rawat** [*1] **Manzil Zaheer** [*2]
Wittawat Jitkrittum [1] **Veeranjaneyulu Sadhanala** [1] **Sadeep Jayasumana** [1] **Aditya Krishna Menon** [1]
Rob Fergus [2] **Sanjiv Kumar** [1]

Abstract

Modern information retrieval (IR) systems consists of multiple stages like retrieval and ranking, with Transformer-based models achieving state-of-the-art performance at each stage. In this paper, we challenge the tradition of using separate models for different stages and ask if a single Transformer encoder can provide relevance score needed in each stage. We present USTAD – a new unified approach to train a single network that can provide powerful ranking scores as a cross-encoder (CE) model as well as factorized embeddings for large-scale retrieval as a dual-encoder (DE) model. Empirically, we find a single USTAD model to be competitive to separate ranking CE and retrieval DE models. Furthermore, USTAD combines well with a novel embedding matching-based distillation, significantly improving CE to DE distillation. It further motivates novel asymmetric architectures for student models to ensure a better embedding alignment between the student and the teacher while ensuring small online inference cost. On standard benchmarks like MS-MARCO, we demonstrate that USTAD with our proposed distillation method leads to asymmetric students with only 1/10th trainable parameter but retaining 95-97% of the teacher performance.

1. Introduction

A typical information retrieval (IR) system comprises two stages: (1) A *retriever* first selects a small subset of potentially relevant candidate documents (out of a large collection) for a given query; and (2) A *reranker* then identifies a precise ranking among the candidates provided by the retriever. *Dual-encoder* (DE) models are the de-facto architecture for retrievers (Lee et al., 2019; Karpukhin et al., 2020a). Such models independently embed queries and documents into a common space via query and document encoders, respectively, and capture their relevance by simple operations on these embeddings such as the inner product. This enables offline creation of a document index, supporting fast retrieval during inference via efficient maximum inner product search implementations (Guo et al., 2020; Johnson et al., 2021), with *online* query embedding generation primarily dictating the inference latency. *Cross-encoder* (CE) models, on the other hand, are preferred as rerankers, owing to their excellent performance (Nogueira & Cho, 2019; Dai & Callan, 2019; Yilmaz et al., 2019). A CE model jointly encodes a query-document pair via a single encoder while enabling early interaction among query and document features. Employing a CE model for retrieval is often infeasible, as it would require processing a given query with *every* document in the collection at inference time.

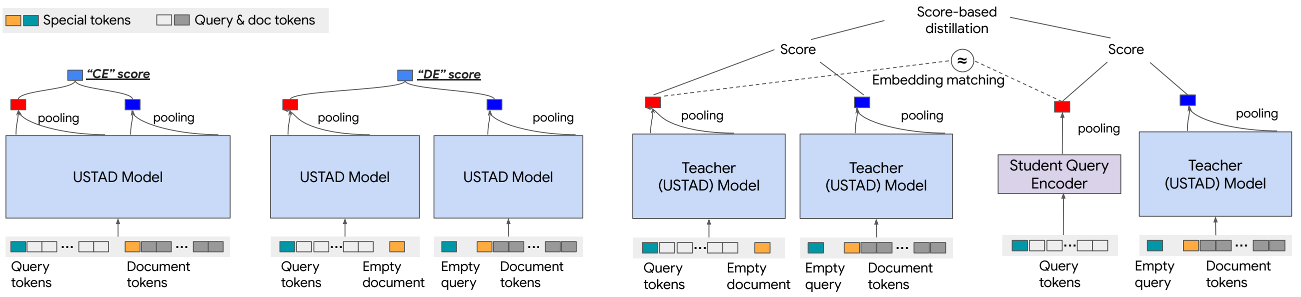
A recent line of work explores *late interaction* models that provide a middle ground between DE and CE models (Pang et al., 2016; Xiong et al., 2017a; Dai et al., 2018; Hofstätter et al., 2020; Khattab & Zaharia, 2020; Menon et al., 2022; Li et al., 2023). These model ensure computational efficiency by utilizing two encoders similar to DE, while allowing for a *complex* interaction beyond simple inner product for better modeling of the true query-document relevance.

The presence of multiple models with different functionalities and quality-cost operating points increases the complexity of developing and maintaining such IR pipelines. Interestingly, all of these models increasingly rely on Transformers (Vaswani et al., 2017) to design the underlying encoders. Given that Transformers provide an extremely powerful architecture (Yun et al., 2020; Menon et al., 2022) that has the ability to produce high-quality individual representations for query and document features in isolation as well as joint query-document representations from query and document features together, it's natural to ask:

Can we train a unified Transformer encoder that can simultaneously enable CE, DE, and late interaction models when combined with different suitable scoring functions?

[*]Equal contribution [1]Google Research, New York, USA [2]Google DeepMind, New York, USA. Correspondence to: Seungyeon Kim <seungyeonk@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) USTAD

(b) USTAD/DE to DE distillation

Figure 1. **Left:** USTAD model architecture (Sec.4). USTAD can provide diverse query-document relevance scores as well as their individual representation by varying the input to the unified model. **Right:** From USTAD or any DE model, we can employ embedding matching-based distillation (Sec. 5) along with traditional score-based distillation. Additionally, student can be configured as asymmetric DE that inherits a large (non-trainable) document encoder from USTAD/DE teacher combined with a small (trainable) query encoder.

In this paper, we affirmatively answer this question by presenting USTAD – a **u**nified **s**ingle-model **t**raining **a**chieving **d**iverse scores. USTAD greatly simplifies the training and maintenance of the overall IR pipeline. It enables joint training of CE, DE, and late interaction models that share a common Transformer encoder without compromising the performance compared to separately trained models.

However, state-of-the-art performance is only achieved by very large encoders (Ni et al., 2022; Neelakantan et al., 2022), which can be prohibitive for many application. Traditionally, *knowledge distillation* (Bucilă et al., 2006; Hinton et al., 2015) serves as a general strategy to train high-quality models with small inference cost by leveraging models with larger encoders as teacher models. In the IR literature, most existing distillation methods only focus on matching the teacher's query-document relevance score for a give query-document pair (see, e.g., Lu et al., 2020; Hofstätter et al., 2020; Chen et al., 2021; Ren et al., 2021; Santhanam et al., 2021; Izacard & Grave, 2021). In this paper, we show that along with score-matching, teacher-student embedding alignment can further reduce the generalization gap between the teacher and the student. Traditional CE models are unable to provide such embeddings for the alignment during distillation; USTAD, in contrast, offers these embeddings, making it a more effective teacher in IR settings.

Our key contributions are as follows:

- We propose USTAD – a novel method to train a unified model to simultaneously realize CE, DE, and late interaction components of an IR pipeline via a single Transformer encoder (Sec. 4).
- We show that USTAD can act as a better teacher for distillation by going beyond standard score-matching employed with vanilla CE teacher models. It can further enable a novel high-performing student DE model with *asymmetric* configuration where the student model inherits the *frozen* document encoder from the teacher and only trains a small query encoder (Sec. 5).
- We provide a comprehensive empirical evaluation of both USTAD and EmbedDistill (Sec. 6) on two standard IR benchmarks – Natural Questions (Kwiatkowski et al., 2019a) and MSMARCO (Nguyen et al., 2016). We also evaluate EmbedDistill on BEIR benchmark (Thakur et al., 2021) which is used to measure the *zero-shot* performance of an IR model.

2. Related Work

Here, we compare our approach to previous efforts aimed at simplifying and unifying IR stacks. We also discuss our contribution in the context of IR distillation as well as distillation with representation alignments in non-IR settings.

Towards the unification of IR models. Due to the complex nature of the IR stack, prior work has attempt to integrate the training of heterogeneously functioning CE and DE models. One attempt is to train CE and DE models iteratively in multi-staged fashion (Qu et al., 2021; Zhang et al., 2022). Because each stage can leverage improved selection of candidate documents for reranking or retrieving, the multi-staged training provided substantial improvement in the quality of the models albeit with a higher computational cost. On the other side, Ren et al. (2021); Lee et al. (2022) pose joint training objective of retrieval and reranking models, or additionally with other models depending on reranking models (e.g., answer extraction model). These joint training approaches encourage the cooperation of these models (albeit they are still different models), and hence improve the overall quality. USTAD takes a different approach since the joint training objectives are posed on one single unified model instead of separate models. Yadav et al. (2022) instead take a post-hoc approach, factorizing representations of the reranking model to extract partial representations that can be used for the retrieval task. Lastly, with the advent of large-language models (LLMs), instruction-tuned models (Asai et al., 2023; Su et al., 2022) can function as a retriever or a reranker, but they are usually prohibitively expensive to deploy at scale.

Distillation for IR. Traditional distillation techniques have been widely applied in the IR literature, often to distill a teacher CE model to a student DE model (Reimers et al., 2019; Li et al., 2020; Chen et al., 2021). Recently, distillation from a DE model (with complex late interaction) to another DE model (with inner-product scoring) has also been considered (Lin et al., 2021; Hofstätter et al., 2021). As for distilling across different model architectures, Lu et al. (2020); Izacard & Grave (2021) consider distillation from a teacher CE model to a student DE model. Hofstätter et al. (2020) conduct an extensive study of knowledge distillation across a wide-range of model architectures. Most existing distillation schemes for IR rely on only teacher scores; by contrast, we propose a geometric approach that also utilizes the teacher *embeddings*. Many recent efforts (Qu et al., 2021; Ren et al., 2021; Santhanam et al., 2021) show that iterative multi-stage (self-)distillation improves upon single-stage distillation (Qu et al., 2021; Ren et al., 2021; Santhanam et al., 2021). These approaches use a model from the previous stage to obtain labels (Santhanam et al., 2021) as well as mine harder-negatives (Xiong et al., 2021). We only focus on the single-stage distillation in this paper.

Distillation with representation alignments. Outside of the IR context, a few prior works proposed to utilize alignment between hidden layers during distillation (Romero et al., 2014; Sanh et al., 2019; Jiao et al., 2020; Aguilar et al., 2020; Zhang & Ma, 2020). Chen et al. (2022) utilize the representation alignment to re-use teacher's classification layer for image classification. Unlike these works, our work is grounded in a rigorous theoretical understanding of the teacher-student (generalization) gap for IR models (cf. Sec. 5). Furthermore, our work differs from these as it needs to address multiple challenges presented by an IR setting: 1) cross-architecture distillation such as USTAD to DE distillation; 2) partial representation alignment of query or document representations as opposed to aligning for the entire input, i.e., a query-documents pair; and 3) catering representation alignment approach to novel IR setups such as asymmetric DE configuration. To the best of our knowledge, our work is first in the IR literature that goes beyond simply matching scores (or its proxies) for distillation.

3. Background

Let $\mathcal{Q}$ and $\mathcal{D}$ denote the query and document spaces, respectively. An IR model is equivalent to a scorer $s : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$, i.e., it assigns a (relevance) score $s(q, d)$ for a query-document pair $(q, d) \in \mathcal{Q} \times \mathcal{D}$. Ideally, we want to learn a scorer such that $s(q, d) > s(q, d')$ *iff* the document d is more relevant to the query q than document d'. We assume access to n labeled training examples $\mathcal{S}_n = \{(q_i, \mathbf{d}_i, \mathbf{y}_i)\}_{i\in[n]}$. Here, $\mathbf{d}_i = (d_{i,1}, \ldots, d_{i,L}) \in \mathcal{D}^L$, $\forall i \in [n]$, denotes a list of L documents and $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,L}) \in \{0, 1\}^L$ denotes the corresponding labels such that $y_{i,j} = 1$ iff the document $d_{i,j}$ is relevant to the query q_i. Given $\mathcal{S}_n$, we learn an IR model by minimizing

$$R(s; \mathcal{S}_n) := \frac{1}{n} \sum\nolimits_{i\in[n]} \ell\big(s_{q_i,\mathbf{d}_i}, \mathbf{y}_i\big), \tag{1}$$

where $s_{q_i,\mathbf{d}_i} := (s(q_i, d_{1,i}), \ldots, s(q_i, d_{1,L}))$ and $\ell\big(s_{q_i,\mathbf{d}_i}, \mathbf{y}_i\big)$ denotes the loss s incurs on $(q_i, \mathbf{d}_i, \mathbf{y}_i)$. Due to space constraint, we defer concrete choices for the loss function ℓ to Appendix A.

While this learning framework is general enough to work with any IR models, next, we formally introduce three families of Transformer-based IR models that are prevalent in the recent literature.

3.1. Transformer-based IR models

Let query $q = (q^1, \ldots, q^{m_1})$ and document $d = (d^1, \ldots, d^{m_2})$ consist of m_1 and m_2 tokens, respectively. We now discuss how Transformers-based CE, DE, and late interaction models process the (q, d) pair.

Cross-encoder model. Let $p = [q; d]$ be the sequence obtained by concatenating q and d. Further, let $\tilde{p}$ be the sequence obtained by adding special tokens such `[CLS]` and `[SEP]` to p. Given an encoder-only Transformer model Enc, the relevance score for the (q, d) pair is

$$s(q, d) = \langle w, \mathrm{pool}\big(\mathrm{Enc}(\tilde{p})\big)\rangle = \langle w, \mathtt{emb}_{q,d}\rangle, \tag{2}$$

where w is a d-dimensional classification vector, and $\mathrm{pool}(\cdot)$ denotes a pooling operation that transforms the contextualized token embeddings $\mathrm{Enc}(\tilde{p})$ to a joint embedding vector $\mathtt{emb}_{q,d}$. `[CLS]`-pooling is a common operation that simply outputs the embedding of the `[CLS]` token as $\mathtt{emb}_{q,d}$.

Dual-encoder model. Let $\tilde{q}$ and $\tilde{d}$ be the sequences obtained by adding appropriate special tokens to q and d, respectively. A DE model comprises two (encoder-only) Transformers Enc_Q and Enc_D, which we call query and document encoders, respectively.[1] Let $\mathtt{emb}_q = \mathrm{pool}\big(\mathrm{Enc}_Q(\tilde{q})\big)$ and $\mathtt{emb}_d = \mathrm{pool}\big(\mathrm{Enc}_D(\tilde{d})\big)$ denote the query and document embeddings, respectively. Now, one can define $s(q, d) = \langle \mathtt{emb}_q, \mathtt{emb}_d\rangle$ to be the relevance score assigned to the (q, d) pair by the DE model.

Late-interaction model. Similar to DE models, such models also embed queries and documents separately; however, they do not use pooling operations, but instead a non-linear scoring function f to enable relatively complex interaction between query and document token embeddings, i.e.

[1] It is common to employ dual-encoder models where query and document encoders are shared.

$s(q,d) = f\big(\mathrm{Enc}_Q(\tilde{q}), \mathrm{Enc}_D(\tilde{d})\big)$. This non-linearity allows to achieve better accuracy than DE, with multiple variants proposed in the literature: ColBERT (Khattab & Zaharia, 2020) uses sum of max interaction, MatchPyramid (Pang et al., 2016) applies a convolutional network, KNRM (Xiong et al., 2017b) performs kernel-based pooling; and ConvKNRM (Dai et al., 2018) further uses a convolutional network on top of learned token embeddings to produce contextual embeddings.

3.2. Score-based distillation for IR models

Most distillation schemes for IR (e.g., Lu et al., 2020; Hofstätter et al., 2020; Chen et al., 2021) rely on teacher relevance scores. Given a training set $\mathcal{S}_n$ and a teacher with scorer s^{t}, one learns a student with scorer s^{s} by minimizing

$$R(s^{\mathrm{s}}, s^{\mathrm{t}}; \mathcal{S}_n) = \frac{1}{n} \sum\nolimits_{i \in [n]} \ell_{\mathrm{d}}\big(s^{\mathrm{s}}_{q,\mathbf{d}_i}, s^{\mathrm{t}}_{q,\mathbf{d}_i}\big), \tag{3}$$

where ℓ_{d} captures the discrepancy between s^{s} and s^{t}. See Appendix A for common choices for ℓ_{d}.

4. USTAD Architecture

Recall that the main focus of this paper is to an end-to-end IR pipeline by designing a unified single Transformer-based model that can simultaneously enable the functioning of different IR models discussed in Sec. 3.1. Towards this, unifying DE and late interaction models appears to be a manageable goal since both models rely on invoking Transformer encoders twice to generate query and document representations, respectively. On the other hand, requiring a single encoder to act as both CE and DE (or late-interaction) model seems a hopeless task at first thought. Standard CE models jointly encode query-document pairs, making it challenging to extract individual query and document embeddings.

More precisely, since the score of a standard CE model takes the form $s(q,d) = \langle w, \mathtt{emb}^{\mathrm{t}}_{q,d} \rangle$, during training, the joint embeddings $\mathtt{emb}^{\mathrm{t}}_{q,d}$ for relevant and irrelevant (q,d) pairs are encouraged to be aligned with w and $-w$, respectively. As a result, $\mathtt{emb}_{q,d}$ produced by the CE-model does not encode global semantic information useful for similarity search. This further leads a standard CE model to produce degenerate query or document embeddings when asked to act as a query or document encoder. In fact, we notice that even the final query and document token representations lack any semantic structure (see Appendix F.2 for details).

We show that this aforementioned shortcoming of standard CE models can be easily addressed by carefully crafting the input to the Transformer encoder and introducing a different loss term for each functionality we want to cultivate in our unified model. In particular, we obtain CE (reranker), DE (retiever), and late-interaction score with the help of a single Transformer encoder Enc as follows:

- **Reranker mode**: We feed the transformer encoder a concatenation of query and document tokens and obtain a sequence of token embedding vectors h_{qd} of the same length. The final score is computed by first generating query and document embeddings by separately pooling the final token embeddings at query and document token positions at input, respectively, and then employing an inner-product (as illustrated in Fig. 1a):

$$\begin{aligned} h_{qd} &= \mathrm{Enc}([q^1, ..., q^{m_1}, d^1, ..., d^{m_2}]) \\ \mathtt{emb}_q &= \mathrm{pool}(h^1_{qd}, ..., h^{m_1}_{qd}]) \\ \mathtt{emb}_d &= \mathrm{pool}(h^{m_1+1}_{qd}, ..., h^{m_1+m_2}_{qd}]) \\ s_{\mathrm{CE}}(q,d) &= \langle \mathtt{emb}_q, \mathtt{emb}_d \rangle \end{aligned} \tag{4}$$

  We refer to the above process of generating separate query and document embeddings in CE mode as *dual-pooling*.

- **Retriever mode**: We independently feed the query and document tokens to the same transformer encoder Enc. The final DE score is obtained by simple inner product between the corresponding pooled representations.

$$\begin{aligned} h_q &= \mathrm{Enc}([q^1, ..., q^{m_1}, 0, ..., 0]) \\ \mathtt{emb}_q &= \mathrm{pool}(h^1_q, ..., h^{m_1}_q]) \\ h_d &= \mathrm{Enc}([0, ..., 0, d^1, ..., d^{m_2}]) \\ \mathtt{emb}_d &= \mathrm{pool}(h^{m_1+1}_d, ..., h^{m_1+m_2}_d]) \\ s_{\mathrm{DE}}(q,d) &= \langle \mathtt{emb}_q, \mathtt{emb}_d \rangle \end{aligned} \tag{5}$$

- **Late interaction mode**: Similar to the retriever mode, Enc independently operates on query and document tokens in this mode. Subsequently, the final score is obtained by a more complex interaction function f operating directly on the encoder-produced token embeddings.

$$\begin{aligned} h_q &= \mathrm{Enc}([q^1, ..., q^{m_1}, 0, ..., 0]) \\ h_d &= \mathrm{Enc}([0, ..., 0, d^1, ..., d^{m_2}]) \\ s_{\mathrm{LI}}(q,d) &= f(h_q, h_d) \end{aligned} \tag{6}$$

  In the work, we focus on the following interaction function introduce in ColBERT (Khattab & Zaharia, 2020):

$$s_{\mathrm{LI}}(q,d) = \sum_{i=1}^{m_1} \max_{1 \le j \le m_2} \langle h^i_q, h^j_d \rangle \tag{7}$$

Training We jointly train various modes of the transformer encoder on a training data $\mathcal{S}_n$ while also aligning the scores produced in different modes via score-matching among the modes. For example, assuming that we are training USTAD to enable CE and DE modes, the overall training objective takes the form:

$$\mathcal{L} = R(s_{\mathrm{CE}}; \mathcal{S}_n) + R(s_{\mathrm{DE}}; \mathcal{S}_n) + \lambda \| s_{\mathrm{CE}} - s_{\mathrm{DE}} \|, \tag{8}$$

where R is defined as in Eq. (1). We can naturally extend the above objective to include the loss and the score-matching

Objective			MRR@10		
CE	**ColBERT**	**DE**	**CE**	**Colbert**	**DE**
✓			40.22	13.09	01.84
	✓		01.54	39.46	03.94
		✓	03.39	32.73	37.37
✓	✓		40.11	39.60	14.40
✓		✓	39.86	33.25	37.44
	✓	✓	03.57	38.99	37.31
✓	✓	✓	40.02	39.06	37.27

Table 1. Ablation of USTAD model for combination of training objectives on MSMARCO dev set.

terms for late interaction modes as well. Above can be easily generalized to handle all three modes. The losses above can be coupled with common techniques to train IR models like hard negative mining (Qu et al., 2021). Furthermore, instead of one-hot data, we can also distill from a teacher model by replace R from Eq. (3). In fact, the score-matching in Eq. (8) can be viewed as performing self-distillation from the stronger (say CE) mode to the weaker (say DE) mode.

Empirically, we find USTAD can learn the diverse scores using a single model. As illustrated in Table 3-5, USTAD matches performance of IR stack (retriever+reranker) with >2x more parameters on NQ benchmark. Additionally, we present ablations on combinations of various scoring modes on MSMARCO benchmark in Table 1.

5. Improved Distillation via USTAD

High performing IR models are often utilized as teacher to train lightweight models for applications with strict latency requirements. Traditionally, teacher provides supervision in terms of query-document relevance score only, but in this section we show the benefit of having access to individual embeddings for query and document from the teacher model. In order to show this benefit, we begin by theoretically analyzing the teacher-student generalization gap when teacher provides such embeddings. Subsequently, informed by our analysis, we identify two novel ways to improve the student model's performance by leveraging a USTAD model.

5.1. Teacher-student generalization gap in IR models

Let $R(s) = \mathbb{E}\left[\ell\left(s_{q,\mathbf{d}}, \mathbf{y}\right)\right]$ be the population version of the empirical risk in Eq. (1), which measures the test time performance of the IR model defined by the scorer s. Thus, $R(s^{\mathrm{s}}) - R(s^{\mathrm{t}})$ denotes the *teacher-student generalization gap*. In the following result, we bound this quantity (see Appendix B.1 for a formal statement and proof).

Theorem 5.1 (Teacher-student generalization gap (informal))**.** *Let $\mathcal{F}$ and $\mathcal{G}$ denote the function classes for the query and document encoders for the student model, respectively. Suppose both the score-based distillation loss ℓ_{d} in Eq. (3) and one-hot (label-dependent) loss ℓ in Eq. (1) is based on binary cross entropy loss (Eq. (17) and (15) in Appendix A respectively). Further, assume that all encoders have the same output dimension and embeddings have their ℓ_2-norm bounded by K. Then, we have*

$$\begin{aligned} &R(s^{\mathrm{s}}) - R(s^{\mathrm{t}}) \qquad (9)\\ &\leq \hat{R}(s^{\mathrm{t}}, \mathcal{S}_n) + \Delta(s^{\mathrm{t}}; \mathcal{S}_n) + K^2\big(\mathbb{E}\left[\left|\sigma(s^{\mathrm{t}}_{q,d}) - y\right|\right]\\ &+ \mathcal{E}_n(\mathcal{F}, \mathcal{G}) + 2KR_{\mathrm{Emb},Q}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n) + 2KR_{\mathrm{Emb},D}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)\end{aligned}$$

where we define $\hat{R}(s^{\mathrm{t}}, \mathcal{S}_n) := 1/n \sum_{i \in [n]} \left|\sigma(s^{\mathrm{t}}_{q_i,d_i}) - y_i\right|$, $\mathcal{E}_n(\mathcal{F}, \mathcal{G}) := \sup_{s^{\mathrm{s}} \in \mathcal{F} \times \mathcal{G}} \left|R(s^{\mathrm{s}}, s^{\mathrm{t}}; \mathcal{S}_n) - \mathbb{E}\ell_{\mathrm{d}}\left(s^{\mathrm{s}}_{q,d}, s^{\mathrm{t}}_{q,d}\right)\right|$; σ denotes the sigmoid function; and $\Delta(s^{\mathrm{t}}; \mathcal{S}_n)$ denotes the deviation between the empirical risk (on $\mathcal{S}_n$) and population risk of the teacher s^{t}. Here, $R_{\mathrm{Emb},Q}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)$ and $R_{\mathrm{Emb},D}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)$ measure misalignment between teacher and student embeddings by focusing on queries and documents, respectively (cf. Eq. (10) & (11) below).

The first three quantities in the bound in Thm. 5.1, namely $\hat{R}(s^{\mathrm{t}}, \mathcal{S}_n)$, $\Delta(s^{\mathrm{t}}; \mathcal{S}_n)$, and $\mathbb{E}[|\sigma(s^{\mathrm{t}}_{q,d}) - y|]$, are *independent* of the student model. These terms solely depend on the quality of the teacher model s^{t}. That said, the teacher-student gap can be made small by reducing the following three terms: 1) uniform deviation of the student's empirical distillation risk from its population version $\mathcal{E}_n(\mathcal{F}, \mathcal{G})$; 2) misalignment between teacher student query embeddings $R_{\mathrm{Emb},Q}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)$; and 3) misalignment between teacher student document embeddings $R_{\mathrm{Emb},D}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)$.

5.2. Two proposed solutions

Embedding matching during distillation. The last two terms in the RHS of Eq. (9) motivate us to propose an *embedding matching*-based distillation, namely EmbedDistill, that explicitly aims to minimize these terms during student training: Given a (q, d) pair, let $\mathtt{emb}^{\mathrm{t}}_q$ and $\mathtt{emb}^{\mathrm{t}}_d$ be the query and document embeddings produced by the query encoder $\mathrm{Enc}^{\mathrm{t}}_Q$ and document encoder $\mathrm{Enc}^{\mathrm{t}}_D$ of the teacher DE model, respectively.[2] Similarly, let $\mathtt{emb}^{\mathrm{s}}_q$ and $\mathtt{emb}^{\mathrm{s}}_d$ denote the query and document embeddings produced by a student DE model with $(\mathrm{Enc}^{\mathrm{s}}_Q, \mathrm{Enc}^{\mathrm{s}}_D)$ as its query and document encoders. Now, EmbedDistill optimizes the following embedding alignment losses in addition to the score-matching loss from Sec. 3.2 to align query and document embeddings of the teacher and student:

$$R_{\mathrm{Emb},Q}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n) = \frac{1}{n} \sum_{q \in \mathcal{S}_n} \left\|\mathtt{emb}^{\mathrm{t}}_q - \mathrm{proj}\big(\mathtt{emb}^{\mathrm{s}}_q\big)\right\| \qquad (10)$$

$$R_{\mathrm{Emb},D}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n) = \frac{1}{n} \sum_{d \in \mathcal{S}_n} \left\|\mathtt{emb}^{\mathrm{t}}_d - \mathrm{proj}\big(\mathtt{emb}^{\mathrm{s}}_d\big)\right\| \qquad (11)$$

[2] It possible to have a single encoder, i.e., $\mathrm{Enc}^{\mathrm{t}}_Q = \mathrm{Enc}^{\mathrm{t}}_D$.

where proj is a projection layer to match student and teacher embedding dimensions. To further help align the embeddings spaces of the teacher and student, we propose to generate similar queries or documents that can naturally help enforce such an alignment globally on the task-specific manifold: Given a set of unlabeled (generated) task-specific query and document pairs $\mathcal{U}_m$, we can further add the embedding matching losses $R_{\mathrm{Emb,Q}}(\mathrm{t}, \mathrm{s}; \mathcal{U}_m)$ or $R_{\mathrm{Emb,D}}(\mathrm{t}, \mathrm{s}; \mathcal{U}_m)$ to our training objective. Please refer to Appendix C for details about task-specific data generation.

Asymmetric DE student. To reduce the $\mathcal{E}_n(\mathcal{F}, \mathcal{G})$ term, we also propose a novel student DE configuration where the student employs the teacher's document encoder (i.e., $\mathrm{Enc}^{\mathrm{s}}_D = \mathrm{Enc}^{\mathrm{t}}_D$) and only train its query encoder, which is much smaller compared to the teacher's query encoder. This not only reduces trainable parameters in the student model, which leads to reduction in $\mathcal{E}_n(\mathcal{F}, \mathcal{G})$ as formalized below, but also immediately makes the $R_{\mathrm{Emb},D}(\mathrm{t}, \mathrm{s}; \mathcal{S}_n)$ term go to 0. For such a setting, it is natural to only employ the embedding matching loss in Eq. (10) as the document embeddings are aligned by design (cf. Fig. 1b).

Proposition 5.2. *Let ℓ_{d} be a distillation loss which is $L_{\ell_{\mathrm{d}}}$-Lipschitz in its first argument. Let $\mathcal{F}$ and $\mathcal{G}$ denote the function classes for the query and document encoders, respectively. Further assume that, for each query and document encoder in our function class, the query and document embeddings have their ℓ_2-norm bounded by K. Then,*

$$\mathcal{E}_n(\mathcal{F}, \mathcal{G}) \leq \mathbb{E}_{\mathcal{S}_n} \frac{48KL_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log\big(N(u, \mathcal{F})N(u, \mathcal{G})\big)}\, du. \tag{12}$$

Furthermore, with a fixed document encoder, i.e., $\mathcal{G} = \{g^*\}$,

$$\mathcal{E}_n(\mathcal{F}, \{g*\}) \leq \mathbb{E}_{\mathcal{S}_n} \frac{48KL_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u, \mathcal{F})}\, du. \tag{13}$$

Here, $N(u, \cdot)$ is the u-covering number of a function class.

Note that Eq. (12) and Eq. (13) correspond to uniform deviation when we train *without* and *with* a frozen document encoder, respectively. It is clear that the bound in Eq. (13) is less than or equal to that in Eq. (12) (because $N(u, \mathcal{G}) \geq 1$ for any u), which alludes to desirable impact of employing a frozen document encoder.

Note that this asymmetric student DE does not incur an increase in latency despite the use of a large teacher document encoder. This is because the large document encoder from USTAD is only needed to create a good quality document index offline, and only the query encoder is evaluated at inference time. Also, the similarity search cost is not increased as the projection layer ensures the same small embedding dimension as in the symmetric DE student. Thus, backed by theoretical reasoning as well as empirical result, we prescribe the asymmetric DE configuration universally.

6. Experiments

We already showcased the efficacy of USTAD framework for producing a unified IR model in Sec. 4. Here, besides providing additional details regarding the empirical results in Sec. 4, we demonstrate the utility EmbedDistill (cf. Sec. 5) in isolation as well as in tandem with USTAD. We also showcase the benefits of combining our distillation approach with query generation methods.

6.1. Setup

Benchmarks and evaluation metrics. We consider two popular IR benchmarks — Natural Questions (NQ) (Kwiatkowski et al., 2019b) and MSMARCO (Nguyen et al., 2016), which focus on finding the most relevant passage/document given a question and a search query, respectively. NQ provides both standard test and dev sets, whereas MSMARCO provides only the dev set that are widely used for common benchmarks. In what follows, we use the terms query (document) and question (passages) interchangeably.

For NQ, we use the *strict* recall as well as the *relaxed* recall metric (Karpukhin et al., 2020a) to evaluate both reranking and retrieval performance. For the reranking, we utilize the top 100 candidates retrieved from AR2-g (Zhang et al., 2022), one of the state-of-the-art (SOTA) models, and also utilize them to construct the reranking training set.

For MSMARCO, we focus on the standard metrics *Mean Reciprocal Rank* (MRR)@10, and *normalized Discounted Cumulative Gain* (nDCG)@10 to evaluate both reranking and retrieval performance. For reranking evaluations, we restrict to reranking only the top 1000 candidate document provided as part of the dataset to be fair, while some works use stronger methods to find better top 1000 candidates for reranking (resulting in higher evaluation numbers).

See Appendix D for details on these evaluation metrics.

Model architectures. USTAD model is based on the pre-trained BERT-base model (Devlin et al., 2019) (12-layer, 768 dim, 110M parameters) and additionally trained for 20k steps with the USTAD training objectives in Eq. (8). We utilized various sizes of DE models as students, based on DistilBERT (Sanh et al., 2019) (6-layer, 768 dim, 67.5M parameters – $\sim$ 2/3 of BERT-base) or BERT-mini (Turc et al., 2019) (4-layer, 256 dim, 11.3M parameters – $\sim$ 1/10 of BERT-base).

For query generation (Appendix C), we employ BART-base (Lewis et al., 2020), an encoder-decoder model, to generate similar questions from each training example's input question (query). During generation, we randomly mask 10% of tokens and inject zero mean Gaussian noise with $\sigma = \{0.1, 0.2\}$ between the encoder and decoder. As for other training-related hyperparameters, see Appendix E.1.

Dataset	**Natural Questions (Dev)**						**MSMARCO (Dev)**			
Method	**67.5M**			**11.3M**			**67.5M**		**11.3M**	
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**MRR@10**	**nDCG@10**	**MRR@10**	**nDCG@10**
Train student directly	39.5	66.4	74.7	34.1	59.8	68.6	27.0	32.2	23.0	29.7
+ Distill from teacher	42.4	70.4	78.1	36.4	62.3	71.5	33.2	38.7	28.6	33.6
+ Inherit doc embeddings	42.8	70.3	78.7	35.4	61.6	71.3	35.4	41.0	30.2	35.6
+ Query embedding matching	43.4	71.3	79.5	37.9	64.8	73.6	36.1	41.7	31.7	37.1
+ Query generation	**43.8**	**71.5**	**80.1**	37.5	64.0	73.3	36.3	42.0	32.1	37.6
Train student using only embedding matching and inherit doc embeddings	43.3	71.3	79.7	36.8	63.8	73.3	**36.9**	**42.6**	34.7	40.4
+ Query generation	43.4	71.5	79.9	**38.1**	**65.2**	**74.7**	36.8	42.5	**35.1**	**40.8**

Table 2. Reranking performance of various student DE models on NQ and MSMARCO dev set, including symmetric DE model (67.5M or 11.3M transformer as both encoders) and asymmetric DE student model (67.5M or 11.3M transformer as query encoder and document embeddings inherited from USTAD teacher). **The USTAD teacher achieves R@1 = 47.4, R@5 = 77.2, R@10 = 83.7, on NQ and MRR@10 = 40.0, nDCG@10 = 45.8 on MSMARCO.**

Method	**#Params**	**R@5**	**R@20**
GAR$^+$-BART (Mao et al., 2021)	406M	73.5	82.2
GAR$^+$-RIDER (Mao et al., 2021)	110M	75.2	83.2
R2-D2 (Fajcik et al., 2021)	110M	76.8	84.5
YONO (Lee et al., 2022)	220M	79.1	86.7
AR2-d (Zhang et al., 2022)	330M	81.5	-
USTAD CE mode	110M	79.9	87.3
EmbedDistill trained DistillBERT	68M	75.0	85.7
EmbedDistill trained BERT-mini	11M	70.8	84.1

Table 3. Reranking performance of our models and other contemporary reranking models on NQ test set. Per standard procedure, the performance is measured by *relaxed recall* metric.

6.2. USTAD to DE distillation

As discussed in Sec. 4 and 5, USTAD provides two benefits: 1) simplified IR setup with a single retrieval and reranking model; and 2) improved distillation via embedding matching. Here, we focus on demonstrating the latter benefit on NQ and MSMARCO benchmarks.

Teacher USTAD model training. For NQ, we utilize AR2-g (Zhang et al., 2022) to collect negatives and provides scores to form a reranking dataset similar to Fajcik et al. (2021): Each question is associated with a single ground truth passage provided in the original dataset and extra 99 negative candidates supplied by retrieval using AR2-g. Given such a dataset, we trained USTAD model from BERT-base initialization, with ListMLE loss (Xia et al., 2008) defined over one positive and 19 randomly sampled negatives from 99 candidates. For MSMARCO, we similarly collected negatives and distilled from SimLM `[CLS]`-pooled CE model[3].

Student DE model training. We consider two kinds of configurations for the student DE model: (1) *Symmetric*: We use identical question and document encoders. We evaluate both DistilBERT and BERT-mini as encoders in the student. (2) *Asymmetric*: The student simply inherits document embeddings from the teacher USTAD model which are *not* updated during the distillation. For trainable query encoder, we use DistilBERT or BERT-mini which are smaller than document encoder.

We evaluate student DE models on various training combinations (i) one-hot loss (cf. Eq. (14) in Appendix A) on training data; (ii) standard distillation loss in (cf. Eq. (16) in Appendix A); and (iii) embedding matching loss in Eq. (10). We used `[CLS]`-pooling for all student encoders. Unlike DPR (Karpukhin et al., 2020a) or AR2, we do not use hard negatives from BM25 or other models, which greatly simplifies our distillation procedure.

Results and discussion. To understand the impact of various proposed configurations and losses, we train models by sequentially adding components and evaluate their reranking performance on NQ and MSMARCO dev set as shown in Table 2, and NQ test set as shown in Table 3.

We begin by directly training a symmetric DE from the dataset without distillation. As expected, moving to distillation brings in considerable gains. Next, we swap the student document encoder with non-trainable document embeddings from the teacher (“Inherit doc embeddings”). This leads to considerable gain in the performance, as per our discussion in Sec. 5.1 on significantly decreasing teacher-student gap via inheriting teacher document embeddings. Now we can introduce EmbedDistill with Eq. (10) for aligning query representations between student and teacher. This improves performance significantly, e.g., it provides ∼1-2 points increase in recall@1, 5, 10 on NQ with students based on DistilBERT and BERT-mini, respectively (Table 2). The excellent performance of distillation to an asymmetric DE

[3] https://github.com/microsoft/unilm/tree/master/simlm to USTAD model via standard score-based distillation (cf. Sec. 3.2)

Dataset	**Natural Questions (Dev)**						**MSMARCO (Dev)**			
Method	**67.5M**			**11.3M**			**67.5M**		**11.3M**	
	R@5	**R@20**	**R@100**	**R@5**	**R@20**	**R@100**	**MRR@10**	**nDCG@10**	**MRR@10**	**nDCG@10**
Train student directly	36.2	59.7	80.0	24.8	44.7	67.5	22.6	27.2	18.6	22.5
+ Distill from teacher	65.3	81.6	91.2	44.3	64.9	81.0	35.0	41.3	28.6	34.1
+ Inherit doc embeddings	69.9	83.9	92.3	56.3	70.9	82.5	35.7	42.2	30.3	36.2
+ Query embedding matching	72.7	86.5	93.9	61.2	75.2	85.1	37.1	43.8	**35.4**	**41.9**
+ Query generation	**73.4**	**86.3**	**93.8**	**64.3**	**77.8**	**87.9**	**37.2**	**43.8**	34.8	41.2
Train student using only embedding matching and inherit doc embeddings	71.4	84.9	92.6	50.2	64.6	76.8	36.6	43.3	31.4	37.6
+ Query generation	71.8	85.0	93.0	54.2	68.9	80.8	36.7	43.4	32.8	39.2

Table 4. Retrieval performance (full recall against all documents in the corpus) of various student DE models on NQ and MSMARCO dev set, including symmetric DE model (67.5M or 11.3M transformer as both encoders) and asymmetric DE student model. **Teacher achieved R@5 = 72.3, R@20 = 86.1, and R@100 = 93.6 on NQ and MRR@10 = 37.2 and nDCG@10 = 44.2 on MSMARCO.**

Method	**#Params**	**R@20**	**R@100**
DPR (Karpukhin et al., 2020a)	220M	78.4	85.4
R2D2 (Fajcik et al., 2021)	220M	80.6	86.7
ACNE (Xiong et al., 2021)	220M	81.9	87.5
RocketQA (Qu et al., 2021)	220M	82.7	88.5
DPR + PAQ (Oğuz et al., 2021)	220M	84.0	89.2
DPR + PAQ (Oğuz et al., 2021)	660M	84.7	89.2
YONO (Lee et al., 2022)	165M	85.2	90.2
AR2-g (Zhang et al., 2022)	220M	85.4	90.0
USTAD DE mode	110M	85.3	89.9
EmbedDistill trained DistilBERT	68M	85.1	89.8
EmbedDistill trained BERT-mini	11M	81.2	87.4

Table 5. Retrieval performance of the proposed method for DE to DE distillation on NQ test set. Per standard procedure, the performance is measured by *relaxed recall* metric.

Method	**#Layers**	**nDCG@10**	**R@100**
SentenceBERT (Reimers et al., 2019)	12	45.7	65.1
DPR (Karpukhin et al., 2020b)	12	22.5	47.7
ANCE (Xiong et al., 2021)	12	40.5	60.0
Contriever (Izacard et al., 2021)	12	46.6	67.0
Jina (Günther et al., 2023)	12	44.5	–
TAS-B (Hofstätter et al., 2021)	6	42.8	64.8
GenQ (Thakur et al., 2021)	6	42.5	64.2
(Wang & Lyu, 2023)	6	40.6	–
EmbedDistill trained DistilBERT	6	44.0	63.5

Table 6. Average BEIR performance of our 6-layer student model trained with EmbedDistill and other competitive baselines and their numbers of trainable parameters. Models are trained on MSMARCO and evaluated on 14 other datasets (the average does not include MSMARCO). – means that paper does not report the number. The full table is at Appendix E.2.

model not only showcases the power of embedding alignment but also highlights the effectiveness of USTAD teacher providing transferable representations.

On top of the two losses (standard distillation and embedding matching), we also use $R_{\mathrm{Emb,Q}}(\mathrm{t}, \mathrm{s}; \mathcal{Q}')$ on additional questions or queries generated from BART (cf. Appendix C). This leads to additional gains in most cases. Furthermore, query generation also proves valuable in another variant of distillation where we eliminate the standard distillation loss and only employ the embedding matching loss in Eq. (10) along with inheriting teacher's document embeddings.

In Table 3, we select the best model (based on the NQ dev set performance) and evaluate it on NQ test set using the standard *relaxed recall* metric. We find that our smaller models (67.5M or 11.3M parameters) are competitive to 10x or 40x larger SOTA models.

6.3. DE to DE distillation

Next, to establish the value of EmbedDistill as a standalone contribution (independent of USTAD), we explore distilling SOTA DE models to smaller DE models via EmbedDistill.

Teacher DE models. We employ AR2 (Zhang et al., 2022)[4] and SentenceBERT-v5 (Reimers et al., 2019)[5] as teacher DE models for NQ and MSMARCO, respectively. Note that both models are based on BERT-base.

Student DE model training. Simliar to Sec. 6.2, we explore two models sizes: DistilBERT and BERT-mini and also consider both *symmetric* and *asymmetric* setups.

Results and discussion. Similar to Sec. 6.2, we sequentially add training methods one by one to understand the impact of each method in Table 4. Compared to the initial direct training, standard distillation provides substantial gains. When we inherit the document encoder from the teacher, followed with the EmbedDistill and query generation, we can observe considerable gain similar to what we observed

[4] https://github.com/microsoft/AR2/tree/main/AR2

[5] https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5

for USTAD to DE distillation in Sec. 6.2.

Finally, we take our best student models, i.e., ones trained using with additional embedding matching loss and using data augmentation from query generation, and evaluate on test sets.

We compare with various prior work and note that most prior work used considerably bigger models in terms of parameters, with larger depth (12 or 24 layers) and width (up to 1024 dims). For NQ, test set results are reported in Table 5. Since MSMARCO does not have any public test set, we instead present results for the BEIR benchmark in Table 6. Please see Table 7 (nDCG@10) and Table 8 (Recall@100) in Appendix E.2 for the detailed results. For both NQ and BEIR, our approach obtains competitive student models as even with 50% fewer parameters (i.e., with 6 layers) our student models can attain $\sim$ 98-99% of teacher's performance. Furthermore, even with 1/10th size of the query encoder, our proposal can achieve 95-97% of teacher's performance.

7. Conclusion

This work presents USTAD, a novel approach that unifies the traditionally separate tasks of retrieval and reranking within a single Transformer model. USTAD offers two significant advantages: (i) a unified architecture for both retrieval and reranking, and (ii) a highly effective distillation technique, EmbedDistill, with asymmetric dual-encoder configuration, achieving competitive performance with a 10x smaller query encoder.

For future work, we plan to extend USTAD to incorporate late interaction scoring (e.g. ColBERT (Khattab & Zaharia, 2020)), which could further enhance its capabilities and versatility in information retrieval tasks. This direction has been partially evaluated in Section 4. Additionally, we aim to explore the integration of this unified formulation in decoder-based retrieval models such as DSI (Tay et al., 2022).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Fan, X., and Guo, C. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7350–7357, 2020.

Asai, A., Schick, T., Lewis, P., Chen, X., Izacard, G., Riedel, S., Hajishirzi, H., and Yih, W.-t. Task-aware retrieval with instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3650–3675, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.225. URL https://aclanthology.org/2023.findings-acl.225.

Bengio, Y. and Senecal, J.-S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4): 713–722, 2008. doi: 10.1109/TNN.2007.912312.

Bousquet, O., Boucheron, S., and Lugosi, G. *Introduction to Statistical Learning Theory*, pp. 169–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_8. URL https://doi.org/10.1007/978-3-540-28650-9_8.

Bucilǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. ACM.

Chen, D., Mei, J.-P., Zhang, H., Wang, C., Feng, Y., and Chen, C. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2022.

Chen, X., He, B., Hui, K., Sun, L., and Sun, Y. Simplified tinybert: Knowledge distillation for document retrieval. In Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., and Sebastiani, F. (eds.), *Advances in Information Retrieval*, pp. 241–248, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72240-1.

Dai, Z. and Callan, J. Deeper text understanding for IR with contextual neural language modeling. In Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., and Scholer, F. (eds.), *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pp. 985–988. ACM, 2019.

Dai, Z., Xiong, C., Callan, J., and Liu, Z. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 126–134, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159659. URL https://doi.org/10.1145/3159652.3159659.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Fajcik, M., Docekal, M., Ondrej, K., and Smrz, P. R2-D2: A Modular Baseline for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 854–870, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.73.

Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., and Kumar, S. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/1908.10396.

Günther, M., Milliken, L., Geuter, J., Mastrapas, G., Wang, B., and Xiao, H. Jina embeddings: A novel set of high-performance sentence embedding models, 2023.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015.

Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M., and Hanbury, A. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666, 2020. URL https://arxiv.org/abs/2010.02666.

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

Izacard, G. and Grave, E. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NTEz-6wysdb.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020a. Association for Computational Linguistics.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020b.

Khattab, O. and Zaharia, M. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*, pp. 39–48. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450380164.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019a. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019b.

Ledoux, M. and Talagrand, M. *Probability in Banach spaces*. Springer-Verlag, 1991.

Lee, H., Kedia, A., Lee, J., Paranjape, A., Manning, C., and Woo, K.-G. You only need one model for open-domain question answering. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pp. 3047–3060, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.198. URL https://aclanthology.org/2022.emnlp-main.198.

Lee, K., Chang, M., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering. In Korhonen, A., Traum, D. R., and Màrquez, L. (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6086–6096. Association for Computational Linguistics, 2019.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Li, C., Yates, A., MacAvaney, S., He, B., and Sun, Y. Parade: Passage representation aggregation for document reranking. *arXiv preprint arXiv:2008.09093*, 2020.

Li, M., Lin, S.-C., Ma, X., and Lin, J. Slim: Sparsified late interaction for multi-vector retrieval with inverted indexes. *arXiv preprint arXiv:2302.06587*, 2023.

Lin, S.-C., Yang, J.-H., and Lin, J. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 163–173, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.repl4nlp-1.17. URL https://aclanthology.org/2021.repl4nlp-1.17.

Lu, W., Jiao, J., and Zhang, R. Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 2645–2652, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412747. URL https://doi.org/10.1145/3340531.3412747.

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., and Chen, W. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 344–350, 2021.

Menon, A., Jayasumana, S., Rawat, A. S., Kim, S., Reddi, S., and Kumar, S. In defense of dual-encoders for neural ranking. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15376–15400. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/menon22a.html.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. MS MARCO: A human generated machine reading comprehension dataset. In Besold, T. R., Bordes, A., d'Avila Garcez, A. S., and Wayne, G. (eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.-W., and Yang, Y. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.669.

Nogueira, R. and Cho, K. Passage re-ranking with BERT. *CoRR*, abs/1901.04085, 2019. URL http://arxiv.org/abs/1901.04085.

Nogueira, R., Yang, W., Lin, J., and Cho, K. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.

Oğuz, B., Lakhotia, K., Gupta, A., Lewis, P., Karpukhin, V., Piktus, A., Chen, X., Riedel, S., Yih, W.-t., Gupta, S., et al. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*, 2021.

Pang, L., Lan, Y., Guo, J., Xu, J., and Cheng, X. A study of matchpyramid models on ad-hoc retrieval. *arXiv preprint arXiv:1606.04648*, 2016.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 5835–5847. Association for Computational Linguistics, 2021.

Reimers, N., Gurevych, I., and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Ren, R., Qu, Y., Liu, J., Zhao, W. X., She, Q., Wu, H., Wang, H., and Wen, J.-R. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.224. URL https://aclanthology.org/2021.emnlp-main.224.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *CoRR*, abs/2112.01488, 2021.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35: 21831–21843, 2022.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. 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)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Wang, Y. and Lyu, H. Query encoder distillation via embedding alignment is a strong baseline method to boost dense retriever online efficiency. *arXiv preprint arXiv:2306.11550*, 2023.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008.

Xiong, C., Dai, Z., Callan, J., Liu, Z., and Power, R. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 55–64, New York, NY, USA, 2017a. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080809. URL https://doi.org/10.1145/3077136.3080809.

Xiong, C., Dai, Z., Callan, J., Liu, Z., and Power, R. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pp. 55–64, 2017b.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

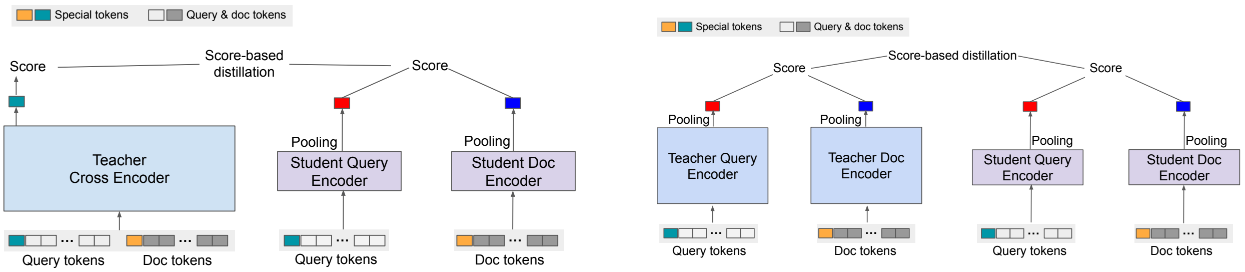
Yadav, N., Monath, N., Angell, R., Zaheer, M., and McCallum, A. Efficient nearest neighbor search for cross-encoder models using matrix factorization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2171–2194, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.140.

Yilmaz, Z. A., Yang, W., Zhang, H., and Lin, J. Cross-domain modeling of sentence-level evidence for document retrieval. 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)*, pp. 3490–3496, Hong Kong, China, November 2019. Association for Computational Linguistics.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Zhang, H., Gong, Y., Shen, Y., Lv, J., Duan, N., and Chen, W. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MR7XubKUFB.

Zhang, L. and Ma, K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020.



(a) Score-based CE to DE distillation

(b) Score-based DE to DE distillation

Figure 2. Illustration of traditional score-based distillation for IR (cf. Section 3.2). Fig. 2a describes distillation from a teacher `[CLS]`-pooled CE model to a student DE model. Fig. 2b depicts distillation from a teacher DE model to a student DE model. Here, both distillation setups employ symmetric DE configurations where query and document encoders of the student model are of the same size.

A. Loss functions

Here, we state various (per-example) loss functions that most commonly define training objectives for IR models. Typically, one hot training with original label is performed using *softmax-based cross-entropy loss* functions:

$$\ell\big(s_{q,\mathbf{d}_i}, \mathbf{y}_i\big) \;=\; -\sum_{j\in[L]} y_{i,j}\cdot\log\left(\frac{\exp(s(q_i,d_{i,j}))}{\sum\limits_{j'\in[L]}\exp(s(q_i,d_{i,j'}))}\right). \tag{14}$$

Alternatively, it is also common to employ a one-vs-all loss function based on *binary cross-entropy loss* as follows:

$$\ell\big(s_{q,\mathbf{d}_i}, \mathbf{y}_i\big) \;=\; -\sum_{j\in[L]}\left(y_{i,j}\cdot\log\left(\frac{1}{1\;+\;\exp(-s(q_i,d_{i,j}))}\right) \;+(1-y_{i,j})\cdot\log\left(\frac{1}{1\;+\;\exp(s(q_i,d_{i,j}))}\right)\right). \tag{15}$$

Note that $\mathbf{d}_i = \{d_{i,j}\}_{j\in[L]}$ can be expanded to include various forms of negatives such as in-batch negatives (Karpukhin et al., 2020b) and sampled negatives (Bengio & Senecal, 2008).

As for distillation (cf. Fig. 2), one can define a distillation objective based on the softmax-based cross-entropy loss as:[6]

$$\ell_{\mathrm{d}}\big(s^{\mathrm{s}}_{q,\mathbf{d}_i}, s^{\mathrm{t}}_{q,\mathbf{d}_i}\big) = -\sum_{j\in[L]}\left(\frac{\exp(s^{\mathrm{t}}_{i,j})}{\sum_{j'\in[L]}\exp(s^{\mathrm{t}}_{i,j'})}\cdot\log\left(\frac{\exp(s^{\mathrm{s}}_{i,j})}{\sum_{j'\in[L]}\exp(s^{\mathrm{s}}_{i,j'})}\right)\right), \tag{16}$$

where $s^{\mathrm{t}}_{i,j} := s^{\mathrm{t}}(q_i, d_{i,j})$ and $s^{\mathrm{s}}_{i,j} := s^{\mathrm{s}}(q_i, d_{i,j})$ denote the teacher and student scores, respectively. On the other hand, the distillation objective with the binary cross-entropy takes the form:

$$\begin{aligned}\ell_{\mathrm{d}}\big(s^{\mathrm{s}}_{q,\mathbf{d}_i}, s^{\mathrm{t}}_{q,\mathbf{d}_i}\big) \;=\; -\sum_{j\in[L]}\Bigg(&\frac{1}{1\;+\;\exp(-s^{\mathrm{t}}_{i,j})}\cdot\log\left(\frac{1}{1\;+\;\exp(-s^{\mathrm{s}}_{i,j})}\right) \;+\\ &\frac{1}{1\;+\;\exp(s^{\mathrm{t}}_{i,j})}\cdot\log\left(\frac{1}{1\;+\;\exp(s^{\mathrm{s}}_{i,j})}\right)\Bigg).\end{aligned} \tag{17}$$

Finally, distillation based on the meas square error (MSE) loss (aka. logit matching) employs the following loss function:

$$\ell_{\mathrm{d}}\big(s^{\mathrm{s}}_{q,\mathbf{d}_i}, s^{\mathrm{t}}_{q,\mathbf{d}_i}\big) = \sum_{j\in[L]}\big(s^{\mathrm{t}}(q_i,d_{i,j}) - s^{\mathrm{s}}(q_i,d_{i,j})\big)^2. \tag{18}$$

[6] It is common to employ temperature scaling with softmax operation. We do not explicitly show the temperature parameter for ease of exposition.

B. Deferred details and proofs from Section 5.1

In this section we present more precise statements and proofs of Theorem 5.1 and Proposition 5.2 (stated informally in Section 5.1 of the main text) along with the necessary background. First, for the ease of exposition, we define new notation which will facilitate theoretical analysis in this section.

Notation. Denote the query and document encoders as $f\colon \mathcal{Q} \to \mathbb{R}^k$ and $g\colon \mathcal{D} \to \mathbb{R}^k$ for the student, and $F\colon \mathcal{Q} \to \mathbb{R}^k, G\colon \mathcal{D} \to \mathbb{R}^k$ for the teacher (in the dual-encoder setting). With q denoting a query and d denoting a document, $f(q)$ and $g(d)$ then denote query and document embeddings, respectively, generated by the student. We define $F(q)$ and $G(d)$ similarly for embeddings by the teacher.[7]

Theorem B.1 (Formal statement of Theorem 5.1)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ denote the function classes for the query and document encoders for the student model, respectively. Given n examples $\mathcal{S}_n = \{(q_i, d_i, y_i)\}_{i\in[n]} \subset \mathcal{Q} \times \mathcal{D} \times \{0,1\}$, let $s^{\mathrm{s}}(q,d) := s^{f,g}(q_i,d_i) = f(q_i)^T g(d_i)$ be the scores assigned to the (q_i,d_i) pair by a dual-encoder model with $f \in \mathcal{F}$ and $g \in \mathcal{G}$ as query and document encoders, respectively. Let ℓ and ℓ_{d} be the binary cross-entropy loss (cf. Eq. (15) with $L = 1$) and the distillation-specific loss based on it (cf. Eq. (17) with $L = 1$), respectively. In particular,*

$$\begin{aligned}
\ell(s^{F,G}(q_i,d_i), y_i) &:= -y_i \log\sigma\left(F(q_i)^\top G(d_i)\right) - (1-y_i)\log\left[1-\sigma\left(F(q_i)^\top G(d_i)\right)\right] \\
\ell_{\mathrm{d}}(s^{f,g}(q_i,d_i), s^{F,G}(q_i,d_i)) &:= -\sigma\left(F(q_i)^\top G(d_i)\right)\cdot\log\sigma\left(f(q_i)^\top g(d_i)\right) \; - \\
&\qquad\qquad \left[1-\sigma\left(F(q_i)^\top G(d_i)\right)\right]\cdot\log\left[1-\sigma\left(f(q_i)^\top g(d_i)\right)\right],
\end{aligned}$$

where σ is the sigmoid function and $s^{\mathrm{t}} := s^{F,G}$ denotes the teacher dual-encoder model with F and Q as its query and document encoders, respectively. Assume that

1. *All encoders $f, g, F,$ and G have the same output dimension.*
2. *$\exists\, K \in (0,\infty)$ such that $\sup_{q\in\mathcal{Q}} \max\{\|f(q)\|_2, \|F(q)\|_2\} \le K$ and $\sup_{d\in\mathcal{D}} \max\{\|g(d)\|_2, \|G(d)\|_2\} \le K$.*

Then, we have

$$\begin{aligned}
\underbrace{\mathbb{E}\left[s^{f,g}(q,d)\right]}_{:=R(s^{\mathrm{s}})=R(s^{f,g})} - \underbrace{\mathbb{E}\left[s^{F,G}(q,d)\right]}_{:=R(s^{\mathrm{t}})=R(s^{F,G})} &\le \underbrace{\sup_{(f,g)\in\mathcal{F}\times\mathcal{G}}\left|R(s^{f,g}, s^{F,G}; \mathcal{S}_n) - \mathbb{E}\left[\ell_{\mathrm{d}}\big(s^{f,g}(q,d), s^{F,G}(q,d)\big)\right]\right|}_{:=\mathcal{E}_n(\mathcal{F},\mathcal{G})} \\
&+ 2K\Big(\underbrace{\frac{1}{n}\sum_{i\in[n]}\|g(d_i)-G(d_i)\|_2}_{:=R_{\mathrm{Emb},D}(\mathrm{t,s};\mathcal{S}_n)} + \underbrace{\frac{1}{n}\sum_{i\in[n]}\|f(q_i)-F(q_i)\|_2}_{:=R_{\mathrm{Emb},Q}(\mathrm{t,s};\mathcal{S}_n)}\Big) + \underbrace{R(s^{F,G};\mathcal{S}_n) - R(s^{F,G})}_{:=\Delta(s^{\mathrm{t}};\mathcal{S}_n)} \\
&+ K^2\Big(\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + \frac{1}{n}\sum_{i\in[n]}\left|\sigma\left(F(q_i)^\top G(d_i)\right) - y_i\right|\Big). \qquad (19)
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
R(s^{f,g}) - R(s^{F,G}) &= R(s^{f,g}) - R(s^{f,g}, s^{F,G}) + R(s^{f,g}, s^{F,G}) - R(s^{F,G}) \\
&\overset{(a)}{\le} K^2\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + R(s^{f,g}, s^{F,G}) - R(s^{F,G}) \\
&= K^2\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + R(s^{f,g}, s^{F,G}) - R(s^{f,g}, s^{F,G}; \mathcal{S}_n) + \\
&\qquad R(s^{f,g}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}) \\
&\overset{(b)}{\le} K^2\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + \mathcal{E}_n(\mathcal{F},\mathcal{G}) + R(s^{f,g}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}) \\
&= K^2\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + \mathcal{E}_n(\mathcal{F},\mathcal{G}) + R(s^{f,g}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}; \mathcal{S}_n) + \\
&\qquad R(s^{F,G}; \mathcal{S}_n) - R(s^{F,G})
\end{aligned}$$

[7] Note that, as per the notations in the main text, we have $(f,g) = (\mathrm{Enc}^{\mathrm{s}}_Q, \mathrm{Enc}^{\mathrm{s}}_D)$ and $(F,G) = (\mathrm{Enc}^{\mathrm{t}}_Q, \mathrm{Enc}^{\mathrm{t}}_D)$. Similarly, we have $(\mathtt{emb}^{\mathrm{t}}_q, \mathtt{emb}^{\mathrm{t}}_d) = (f(q), g(d))$ and $(\mathtt{emb}^{\mathrm{t}}_q, \mathtt{emb}^{\mathrm{t}}_d) = (F(q), G(d))$.

$$
\begin{aligned}
&\overset{(c)}{\leq} K^2\mathbb{E}\left[\left|\sigma(F(q)^\top G(d)) - y\right|\right] + \mathcal{E}_n(\mathcal{F}, \mathcal{G}) + \underbrace{R(s^{F,G}; \mathcal{S}_n) - R(s^{F,G})}_{:=\Delta(s^{\mathrm{t}};\mathcal{S}_n)} + \\
&\quad \frac{2K}{n}\sum_{i\in[n]}\|g(d_i) - G(d_i)\|_2 \ + \frac{2K}{n}\sum_{i\in[n]}\|f(q_i) - F(q_i)\|_2 + \\
&\quad \frac{K^2}{n}\sum_{i\in[n]}\left|\sigma\left(F(q_i)^\top G(d_i)\right) - y_i\right| \qquad (20)
\end{aligned}
$$

where (a) follows from Lemma B.3, (b) follows from the definition of $\mathcal{E}_n(\mathcal{F}, \mathcal{G})$, and (c) follows from Proposition B.2. □

B.1. Bounding the difference between student's empirical *distillation* risk and teacher's empirical risk

Lemma B.2. *Given n examples $\mathcal{S}_n = \{(q_i, d_i, y_i)\}_{i\in[n]} \subset \mathcal{Q} \times \mathcal{D} \times \{0, 1\}$, let $s^{f,g}(q_i, d_i) = f(q_i)^T g(d_i)$ be the scores assigned to the (q_i, d_i) pair by a dual-encoder model with f and g as query and document encoders, respectively. Let ℓ and ℓ_{d} be the binary cross-entropy loss (cf. Eq. (15) with $L = 1$) and the distillation-specific loss based on it (cf. Eq. (17) with $L = 1$), respectively. In particular,*

$$
\begin{aligned}
\ell(s^{F,G}(q_i, d_i), y_i) &:= -y_i \log\sigma\left(F(q_i)^\top G(d_i)\right) - (1 - y_i)\log\left[1 - \sigma\left(F(q_i)^\top G(d_i)\right)\right] \\
\ell_{\mathrm{d}}(s^{f,g}(q_i, d_i), s^{F,G}(q_i, d_i)) &:= -\sigma\left(F(q_i)^\top G(d_i)\right)\cdot\log\sigma\left(f(q_i)^\top g(d_i)\right) \ - \\
&\qquad [1 - \sigma\left(F(q_i)^\top G(d_i)\right)]\cdot\log\left[1 - \sigma\left(f(q_i)^\top g(d_i)\right)\right],
\end{aligned}
$$

where σ is the sigmoid function and $s^{F,G}$ denotes the teacher dual-encoder model with F and Q as its query and document encoders, respectively. Assume that

1. *All encoders f, g, F, and G have the same output dimension $k \geq 1$.*
2. *$\exists\, K \in (0, \infty)$ such that $\sup_{q\in\mathcal{Q}}\max\{\|f(q)\|_2, \|F(q)\|_2\} \leq K$ and $\sup_{d\in\mathcal{D}}\max\{\|g(d)\|_2, \|G(d)\|_2\} \leq K$.*

Then, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]}\ell_{\mathrm{d}}\big(s^{f,g}(q_i, d_i), s^{F,G}(q_i, d_i)\big) - \frac{1}{n}\sum_{i\in[n]}\ell\big(s^{F,G}(q_i, d_i), y_i\big) \leq \\
&\qquad \frac{2K}{n}\sum_{i\in[n]}\|g(d_i) - G(d_i)\|_2 \ + \frac{2K}{n}\sum_{i\in[n]}\|f(q_i) - F(q_i)\|_2 + \\
&\qquad \frac{K^2}{n}\sum_{i\in[n]}\left|\sigma\left(F(q_i)^\top G(d_i)\right) - y_i\right|. \qquad (21)
\end{aligned}
$$

Proof. We first note that the distillation loss can be rewritten as

$$
\ell_{\mathrm{d}}\big(s^{f,g}(q, d), s^{F,G}(q, d)\big) = \big(1 - \sigma(F(q)^\top G(d)\big)\, f(q)^\top g(d) + \gamma(-f(q)^\top g(d)),
$$

where $\gamma(v) := \log[1 + e^v]$ is the softplus function. Similarly, the one-hot (label-dependent) loss can be rewritten as

$$
\ell\big(s^{F,G}(q, d), y\big) = (1 - y)F(q)^\top G(d) + \gamma(-F(q)^\top G(d)).
$$

Recall from our notation in Section 3 that

$$
R(s^{f,g}, s^{F,G}; \mathcal{S}_n) := \frac{1}{n}\sum_{i\in[n]}\ell_{\mathrm{d}}\big(s^{f,g}(q_i, d_i), s^{F,G}(q_i, d_i)\big), \qquad (22)
$$

$$
R(s^{F,G}; \mathcal{S}_n) := \frac{1}{n}\sum_{i\in[n]}\ell\big(s^{F,G}(q_i, d_i), y_i\big), \qquad (23)
$$

as the empirical risk based on the distillation loss, and the empirical risk based on the label-dependent loss, respectively. With this notation, the quantity to upper bound can be rewritten as

$$\begin{aligned} R(s^{f,g}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}; \mathcal{S}_n) = & \underbrace{R(s^{f,g}, , s^{F,G}; \mathcal{S}_n) - R(s^{f,G}, s^{F,G}; \mathcal{S}_n)}_{:=\square_1} + \\ & \underbrace{R(s^{f,G}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}, s^{F,G}; \mathcal{S}_n)}_{:=\square_2} + \underbrace{R(s^{F,G}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}; \mathcal{S}_n)}_{:=\square_3}. \end{aligned} \tag{24}$$

We start by bounding $\square_1$ as

$$\begin{aligned} \square_1 &= \frac{1}{n} \sum_{i \in [n]} \Big(\ell_{\rm d}\big(s^{f,g}(q_i, d_i), s^{F,G}(q_i, d_i)\big) - \ell_{\rm d}\big(s^{f,G}(q_i, d_i), s^{F,G}(q_i, d_i)\big) \Big) \\ &= \frac{1}{n} \sum_{i \in [n]} \Big(\big(1 - \sigma(F(q_i)^\top G(d_i))\big) f(q_i)^\top g(d_i) + \gamma(-f(q_i)^\top g(d_i)) \\ &\qquad\qquad - \big(1 - \sigma(F(q_i)^\top G(d_i))\big) f(q_i)^\top G(d_i) - \gamma(-f(q_i)^\top G(d_i)) \Big) \\ &= \frac{1}{n} \sum_{i \in [n]} \Big(f(q_i)^\top \big(g(d_i) - G(d_i)\big) \big(1 - \sigma(F(q_i)^\top G(d_i))\big) \\ &\qquad\qquad + \gamma(-f(q_i)^\top g(d_i)) - \gamma(-f(q_i)^\top G(d_i)) \Big) \\ &\overset{(a)}{\leq} \frac{1}{n} \sum_{i \in [n]} \Big(f(q_i)^\top \big(g(d_i) - G(d_i)\big) \big(1 - \sigma(F(q_i)^\top G(d_i))\big) + \big| f(q_i)^\top g(d_i) - f(q_i)^\top G(d_i) \big| \Big) \\ &\overset{(b)}{\leq} \frac{1}{n} \sum_{i \in [n]} \Big(\|f(q_i)\| \|g(d_i) - G(d_i)\| \big(1 - \sigma(F(q_i)^\top G(d_i))\big) + \|f(q_i)\| \|g(d_i) - G(d_i)\| \Big) \\ &\leq \frac{K}{n} \sum_{i \in [n]} \|g(d_i) - G(d_i)\|_2 \big(2 - \sigma(F(q_i)^\top G(d_i))\big) \Big) \\ &\leq \frac{2K}{n} \sum_{i \in [n]} \|g(d_i) - G(d_i)\|_2, \end{aligned} \tag{25}$$

where at (a) we use the fact that γ is a Lipschitz continuous function with Lipschitz constant 1, and at (b) we use Cauchy-Schwarz inequality.

Similarly for $\square_2$, we proceed as

$$\begin{aligned} \square_2 &= \frac{1}{n} \sum_{i \in [n]} \Big(\ell_{\rm d}\big(s^{f,G}(q_i, d_i), s^{F,G}(q_i, d_i)\big) - \ell_{\rm d}\big(s^{F,G}(q_i, d_i), s^{F,G}(q_i, d_i)\big) \Big) \\ &= \frac{1}{n} \sum_{i \in [n]} \Big(\big(1 - \sigma(F(q_i)^\top G(d_i))\big) f(q_i)^\top G(d_i) + \gamma(-f(q_i)^\top G(d_i)) \\ &\qquad\qquad - \big(1 - \sigma(F(q_i)^\top G(d_i))\big) F(q_i)^\top G(d_i) - \gamma(-F(q_i)^\top G(d_i)) \Big) \\ &= \frac{1}{n} \sum_{i \in [n]} \Big(G(d_i)^\top (f(q_i) - F(q_i)) \big(1 - \sigma(F(q_i)^\top G(d_i))\big) \\ &\qquad\qquad + \gamma(-f(q_i)^\top G(d_i)) - \gamma(-F(q_i)^\top G(d_i)) \Big) \\ &\leq \frac{1}{n} \sum_{i \in [n]} \Big(\|G(d_i)\| \|f(q_i) - F(q_i)\| + \big| f(q_i)^\top G(d_i) - F(q_i)^\top G(d_i) \big| \Big) \\ &\leq \frac{2K}{n} \sum_{i \in [n]} \|f(q_i) - F(q_i)\|_2. \end{aligned} \tag{26}$$

$\square_3$ can be bounded as

$$\begin{aligned}
\square_3 &= R(s^{F,G}, s^{F,G}; \mathcal{S}_n) - R(s^{F,G}; \mathcal{S}_n) \\
&= \frac{1}{n}\sum_{i\in[n]} \Big(\ell_{\rm d}\big(s^{F,G}(q_i,d_i), s^{F,G}(q_i,d_i)\big) - \ell\big(s^{F,G}(q_i,d_i), y_i\big)\Big) \\
&= \frac{1}{n}\sum_{i\in[n]} \Big(\big(1-\sigma(F(q_i)^\top G(d_i))\big)\, F(q_i)^\top G(d_i) + \gamma(-F(q_i)^\top G(d_i)) \\
&\qquad - (1-y_i)F(q_i)^\top G(d_i) - \gamma(-F(q_i)^\top G(d_i))\Big) \\
&= \frac{1}{n}\sum_{i\in[n]} \Big(\big(1-\sigma(F(q_i)^\top G(d_i)) - (1-y_i)\big) F(q_i)^\top G(d_i)\Big) \\
&\le \frac{K^2}{n}\sum_{i\in[n]} \big|\sigma(F(q_i)^\top G(d_i)) - y_i\big| \,. \qquad (27)
\end{aligned}$$

Combining Eq. 24, 25, 26, and 27 establishes the bound in Eq. 21. □

Lemma B.3. *Given an example $(q,d,y) \in \mathcal{Q}\times\mathcal{D}\times\{0,1\}$, let $s^{f,g}(q,d) = f(q)^T g(d)$ be the scores assigned to the (q,d) pair by a dual-encoder model with f and g as query and document encoders, respectively. Let ℓ and $\ell_{\rm d}$ be the binary cross-entropy loss (cf. Eq. (15) with $L=1$) and the distillation-specific loss based on it (cf. Eq. (17) with $L=1$), respectively. In particular,*

$$\begin{aligned}
\ell(s^{f,g}(q,d), y) &:= -y\log\sigma\left(f(q)^\top g(d)\right) - (1-y)\log\left[1-\sigma\left(f(q)^\top g(d)\right)\right] \\
\ell_{\rm d}(s^{f,g}(q,d), s^{F,G}(q,d)) &:= -\sigma\left(F(q)^\top G(d)\right)\cdot\log\sigma\left(f(q)^\top g(d)\right) - \\
&\qquad [1-\sigma\left(F(q)^\top G(d)\right)]\cdot\log\left[1-\sigma\left(f(q)^\top g(d)\right)\right],
\end{aligned}$$

where σ is the sigmoid function and $s^{F,G}$ denotes the teacher dual-encoder model with F and Q as its query and document encoders, respectively. Assume that

1. *All encoders f, g, F, and G have the same output dimension $k \ge 1$.*
2. *$\exists\, K \in (0,\infty)$ such that $\sup_{q\in\mathcal{Q}}\max\{\|f(q)\|_2, \|F(q)\|_2\} \le K$ and $\sup_{d\in\mathcal{D}}\max\{\|g(d)\|_2, \|G(d)\|_2\} \le K$.*

Then, we have

$$\underbrace{\mathbb{E}\big[\ell\big(s^{f,g}(q,d),y\big)\big]}_{:=R(s^{f,g})} - \underbrace{\mathbb{E}\left[\ell_{\rm d}\big(s^{f,g}(q,d), s^{F,G}(q,d)\big)\right]}_{:=R(s^{f,g}, s^{F,G})} \le\; K_Q K_D \mathbb{E}\left[\big|\sigma(F(q)^\top G(d)) - y\big|\right] \qquad (28)$$

where expectation are defined by a joint distribution $\mathbb{P}(q,d,y)$ over $\mathcal{Q}\times\mathcal{D}\times\{0,1\}$

Proof. Similar to the proof of Proposition B.2, we utilize the fact that

$$\begin{aligned}
\ell\big(s^{F,G}(q,d),y\big) &= (1-y)F(q)^\top G(d) + \gamma(-F(q)^\top G(d)), \\
\ell_{\rm d}\big(s^{f,g}(q,d), s^{F,G}(q,d)\big) &= \big(1-\sigma(F(q)^\top G(d)\big)\, f(q)^\top g(d) + \gamma(-f(q)^\top g(d)),
\end{aligned}$$

where $\gamma(v) := \log[1+e^v]$ is the softplus function. Now,

$$\begin{aligned}
&\mathbb{E}\left[\ell\big(s^{f,g}(q,d),y\big) - \ell_{\rm d}\big(s^{f,g}(q,d), s^{F,G}(q,d)\big)\right] \qquad (29) \\
&\quad= \mathbb{E}\left[(1-y)f(q)^\top g(d) + \gamma(-f(q)^\top g(d))\right] \\
&\qquad - \mathbb{E}\left[\big(1-\sigma(F(q)^\top G(d))\big)\, f(q)^\top g(d) + \gamma(-f(q)^\top g(d))\right] \\
&\quad= \mathbb{E}\left[\Big(1-y-\big(1-\sigma(F(q)^\top G(d))\big)\Big)F(q)^\top G(d)\right] \\
&\quad\le K^2\mathbb{E}\left[\big|\sigma(F(q)^\top G(d)) - y\big|\right], \qquad (30)
\end{aligned}$$

which completes the proof. □

B.2. Uniform deviation bound

Let $\mathcal{F}$ denote the class of functions that map queries in $\mathcal{Q}$ to their embeddings in $\mathbb{R}^k$ via the query encoder. Define $\mathcal{G}$ analogously for the doc encoder, which consists of functions that map documents in $\mathcal{D}$ to their embeddings in $\mathbb{R}^k$. To simplify exposition, we assume that each training example consists of a single relevant or irrelevant document for each query, i.e., $L = 1$ in Section 3. Let

$$\mathcal{F}\mathcal{G} = \{(q, d) \mapsto f(q)^\top g(d) \ \mid \ f \in \mathcal{F}, g \in \mathcal{G}\}$$

Given $\mathcal{S}_n = \{(q_i, d_i, y_i) : i \in [n]\}$, let $N(\epsilon, \mathcal{H})$ denote the ϵ-covering number of a function class $\mathcal{H}$ with respect to $L_2(\mathbb{P}_n)$ norm, where $\|h\|^2_{L_2(\mathbb{P}_n)} := \|h\|_n^2 := \frac{1}{n}\sum_{i=1}^n \|h(q_i, d_i)\|_2^2$. Depending on the context, the functions in $\mathcal{H}$ may map to $\mathbb{R}$ or $\mathbb{R}^d$.

Proposition B.4. *Let s^{t} be scorer of a teacher model and ℓ_{d} be a distillation loss function which is $L_{\ell_{\mathrm{d}}}$-Lipschitz in its first argument. Let the embedding functions in $\mathcal{F}$ and $\mathcal{G}$ output vectors with ℓ_2 norms at most K. Define the uniform deviation*

$$\mathcal{E}_n(\mathcal{F}, \mathcal{G}) = \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left| \frac{1}{n} \sum\nolimits_{i \in [n]} \ell_{\mathrm{d}}\big(f(q_i)^\top g(d_i), s^{\mathrm{t}}_{q_i, d_i}\big) - \mathbb{E}_{q,d} \ell_{\mathrm{d}}\big(f(q)^\top g(d), s^{\mathrm{t}}_{q,d}\big) \right|.$$

For any $g^* \in \mathcal{G}$, we have

$$\mathbb{E}_{\mathcal{S}_n} \mathcal{E}_n(\mathcal{F}, \mathcal{G}) \le \mathbb{E}_{\mathcal{S}_n} \frac{48 K L_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u, \mathcal{F}) + \log N(u, \mathcal{G})}\, du,$$

$$\mathbb{E}_{\mathcal{S}_n} \mathcal{E}_n(\mathcal{F}, \{g^*\}) \le \mathbb{E}_{\mathcal{S}_n} \frac{48 K L_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u, \mathcal{F})}\, du.$$

Proof of Proposition B.4. We first symmetrize excess risk to get Rademacher complexity, then bound the Rademacher complexity with Dudley's entropy integral.

For a training set $\mathcal{S}_n$, the empirical Rademacher complexity of a class of functions $\mathcal{H}$ that maps $\mathcal{Q} \times \mathcal{D}$ to $\mathbb{R}$ is defined by

$$\mathrm{Rad}_n(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(q_i, d_i),$$

where $\{\varepsilon_i\}$ denote i.i.d. Rademacher random variables taking the value in $\{+1, -1\}$ with equal probability. By symmetrization (Bousquet et al., 2004) and the fact that ℓ_{d} is $L_{\ell_{\mathrm{d}}}$-Lipschitz in its first argument, we get

$$E_{\mathcal{S}_n} \mathcal{E}_n(\mathcal{F}, \mathcal{G}) \le 2 L_{\ell_{\mathrm{d}}} \mathbb{E}_{\mathcal{S}_n} \mathrm{Rad}_n(\mathcal{F}\mathcal{G}).$$

Then, Dudley's entropy integral (see, e.g., Ledoux & Talagrand, 1991) gives

$$\mathrm{Rad}_n(\mathcal{F}\mathcal{G}) \le \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u, \mathcal{F}\mathcal{G})}\, du.$$

From Lemma B.5 with $K_Q = K_D = K$, for any $u > 0$,

$$N(u, \mathcal{F}\mathcal{G}) \le N\left(\frac{u}{2K}, \mathcal{F}\right) N\left(\frac{u}{2K}, \mathcal{G}\right).$$

Putting these together,

$$\mathbb{E}_{\mathcal{S}_n} \mathcal{E}_n(\mathcal{F}, \mathcal{G}) \le \frac{24 L_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u/2K, \mathcal{F}) + \log N(u/2K, \mathcal{G})}\, du. \tag{31}$$

Following the same steps with $\mathcal{G}$ replaced by $\{g^*\}$, we get

$$\mathbb{E}_{\mathcal{S}_n} \mathcal{E}_n(\mathcal{F}, \{g^*\}) \le \frac{24 L_{\ell_{\mathrm{d}}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(u/2K, \mathcal{F})}\, du \tag{32}$$

By changing variable in Eq. (31) and Eq. (32), we get the stated bounds. □

For $f : \mathcal{Q} \to \mathbb{R}^k, g : \mathcal{D} \to \mathbb{R}^k$, define $fg : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ by $fg(q, d) = f(q)^\top g(d)$.

Lemma B.5. *Let $f_1, \ldots, f_N$ be an ϵ-cover of $\mathcal{F}$ and $g_1, \ldots, g_M$ be an ϵ-cover of $\mathcal{G}$ in $L_2(\mathbb{P}_n)$ norm. Let $\sup_{f \in \mathcal{F}} \sup_{q \in \mathcal{Q}} \|f(q)\|_2 \leq K_Q$ and $\sup_{g \in \mathcal{G}} \sup_{d \in \mathcal{D}} \|g(d)\|_2 \leq K_D$. Then,*

$$\{f_i g_j \mid i \in [N], j \in [M]\}$$

is a $(K_Q + K_D)\epsilon$-cover of $\mathcal{F}\mathcal{G}$.

Proof of Lemma B.5. For arbitrary $f \in \mathcal{F}, g \in \mathcal{G}$, there exist $\tilde{f} \in \{f_1, \ldots, f_N\}, \tilde{g} \in \{g_1, \ldots, g_M\}$ such that $\|f - \tilde{f}\|_n \leq \epsilon$, $\|g - \tilde{g}\|_n \leq \epsilon$. It is sufficient to show that $\|fg - \tilde{f}\tilde{g}\|_n \leq (K_Q + K_D)\epsilon$. Decomposing using triangle inequality,

$$\begin{aligned} \|fg - \tilde{f}\tilde{g}\|_n &= \|fg - f\tilde{g} + f\tilde{g} - \tilde{f}\tilde{g}\|_n \\ &\leq \|fg - f\tilde{g}\|_n + \|f\tilde{g} - \tilde{f}\tilde{g}\|_n. \end{aligned} \tag{33}$$

To bound the first term, using Cauchy-Schwartz inequality, we can write

$$\frac{1}{n} \sum_{i=1}^{n} \left(f(q_i)^\top g(d_i) - \tilde{f}(q_i)^\top \tilde{g}(d_i) \right)^2 \leq \sup_{q \in \mathcal{Q}} \|f(q)\|_2^2 \cdot \frac{1}{n} \sum_{i=1}^{n} \|(g - \tilde{g})(d_i)\|_2^2.$$

Therefore

$$\|fg - f\tilde{g}\|_n \leq K_Q \|g - \tilde{g}\|_n \leq K_Q \epsilon.$$

Similarly

$$\|f\tilde{g} - \tilde{f}\tilde{g}\|_n \leq K_D \|f - \tilde{f}\|_n \leq K_D \epsilon$$

Plugging these in Eq. (33), we get

$$\|fg - \tilde{f}\tilde{g}\|_n \leq (K_Q + K_D)\epsilon.$$

This completes the proof. □

C. Task-specific data generation (query generation)

Data augmentation as a general technique has been previously considered in the IR literature (see, e.g., Nogueira et al., 2019; Oğuz et al., 2021; Izacard et al., 2021), especially in data-limited, out-of-domain, or zero-shot settings. As EmbedDistill (Section 5) aims to align the embeddings spaces of the teacher and student, the ability to generate similar queries or documents can naturally help enforce such an alignment globally on the task-specific manifold. Given a set of unlabeled task-specific query and document pairs $\mathcal{U}_m$, we can further add the embedding matching losses $R_{\mathrm{Emb,Q}}(\mathrm{t}, \mathrm{s}; \mathcal{U}_m)$ or $R_{\mathrm{Emb,D}}(\mathrm{t}, \mathrm{s}; \mathcal{U}_m)$ to our training objective.

In other words, we introduced task-specific data generation to encourage geometric matching in local regions, which can aid in transferring more knowledge in confusing neighborhoods. As expected, this further improves the distillation effectiveness on top of the embedding matching in most cases.

In the case of inheriting the document encoder, we generate queries from the observed examples by adding local perturbation in the data manifold (embedding space). Specifically, we employ an off-the-shelf encoder-decoder model – BART-base (Lewis et al., 2020). First, we embed an observed query in the corresponding dataset. Second, we add a small perturbation to the query embedding. Finally, we decode the perturbed embedding to generate a new query in the input space. Formally, the generated query x' given an original query x takes the form $x' = \mathrm{Dec}(\mathrm{Enc}(x) + \epsilon)$, where $\mathrm{Enc}()$ and $\mathrm{Dec}()$ correspond to the encoder and the decoder from the off-the-shelf model, respectively, and ϵ is an isotropic Gaussian noise. Furthermore, we also randomly mask the original query tokens with a small probability. We generate two new queries from an observed query and use them as additional data points during our distillation procedure.

As a comparison, we tried adding the same size of random sampled queries instead of the ones generated via the method described above. That did not show any benefit, which justifies the use of our query/question generation method.

D. Evaluation metric details

For NQ, we evaluate models with full *strict* recall metric, meaning that the model is required to find a *golden* passage from the whole set of candidates (21M). Specifically, for $k \geq 1$, recall@k or R@k denotes the percentage of questions for which the associated golden passage is among the k passages that receive the highest relevance scores by the model. In addition, we also present results for *relaxed* recall metric considered by Karpukhin et al. (2020a), where R@k denotes the percentage of questions where the corresponding answer string is present in at least one of the k passages with the highest model (relevance) scores.

For both MSMARCO retrieval and re-ranking tasks, we follow the standard evaluation metrics *Mean Reciprocal Rank*(MRR)@10 and *normalized Discounted Cumulative Gain* (nDCG)@10. For retrieval tasks, these metrics are computed with respect to the whole set of candidates passages (8.8M). On the other hand, for re-ranking task, the metrics are computed with respect to BM25 generated 1000 candidate passages *–the originally provided–* for each query. Please note that some papers use more powerful models (e.g., DE models) to generate the top 1000 candidate passages, which is not a standard re-ranking evaluation and should not be compared directly. We report 100 × MRR@10 and 100 × nDCG@10, as per the convention followed in the prior works.

E. Experimental details and additional results

E.1. Hyperparameters

Optimization. For all of our experiments, we use ADAM weight decay optimizer with a short warm up period (5000 steps) and a linear decay schedule. We use the initial learning rate of 2.8×10^{-5} for training the teacher model, 1×10^{-4} for DE student models from USTAD, and 2.8×10^{-5} for DE to DE experiments. We chose batch sizes to be 128.

Loss weighting. We used 1.0 for the main scoring-based loss and tried 0.02 and 0.005 for embedding matching loss and picked the best performing one.

E.2. Additional results on BEIR benchmark

See Table 7 (NDCG@10) and Table 8 (Recall@100) for BEIR benchmark results. All numbers are from BEIR benchmark paper (Thakur et al., 2021). As common practice, non-public benchmark sets[8], {BioASQ, Signal-1M(RT), TREC-NEWS, Robust04}, are removed from the table. Following the original BEIR paper (Thakur et al., 2021) (Table 9 and Appendix G from the original paper), we utilized Capped Recall@100 for TREC-COVID dataset.

Table 7. In-domain and zero-shot retrieval performance on BEIR benchmark (Thakur et al., 2021), as measured by **nDCG@10**. All the baseline number in the table are taken from (Thakur et al., 2021). We exclude (in-domain) MSMARCO from average computation as per common practice.

	Lexical	Sparse			Dense					
Model and size (→)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**GenQ**	**SentenceBERT** *(our teacher)*	**EmbedDistill** *(ours)*
Dataset (↓)				12-layer	12-layer	12-layer	6-layer	12-layer	12-layer	*6-layer*
MS MARCO	22.8	29.6‡	35.1‡	33.8‡	17.7	38.8‡	40.8‡	40.8‡	47.1‡	46.6‡
TREC-COVID	65.6	40.6	53.8	71.3	33.2	65.4	48.1	61.9	75.4	72.3
NFCorpus	32.5	28.3	30.1	32.8	18.9	23.7	31.9	31.9	31.0	30.7
NQ	32.9	18.8	39.8	39.9	47.4‡	44.6	46.3	35.8	51.5	50.8
HotpotQA	60.3	50.3	49.2	58.0	39.1	45.6	58.4	53.4	58.0	56.0
FiQA-2018	23.6	19.1	19.8	29.1	11.2	29.5	30.0	30.8	31.8	29.5
ArguAna	31.5	30.9	27.9	34.9	17.5	41.5	42.9	49.3	38.5	34.9
Touché-2020	36.7	15.6	17.5	34.7	13.1	24.0	16.2	18.2	22.9	24.7
CQADupStack	29.9	26.8	25.7	32.5	15.3	29.6	31.4	34.7	33.5	30.6
Quora	78.9	69.1	63.0	80.2	24.8	85.2	83.5	83.0	84.2	81.4
DBPedia	31.3	17.7	31.4	33.1	26.3	28.1	38.4	32.8	37.7	35.9
SCIDOCS	15.8	12.4	12.6	16.2	07.7	12.2	14.9	14.3	14.8	14.4
FEVER	75.3	35.3	59.6	71.4	56.2	66.9	70.0	66.9	76.7	76.9
Climate-FEVER	21.3	06.6	08.2	20.1	14.8	19.8	22.8	17.5	23.5	22.5
SciFact	66.5	63.0	58.2	67.5	31.8	50.7	64.3	64.4	59.8	55.5
AVG (w/o MSMARCO)	43.0	31.0	35.5	44.4	25.5	40.5	42.8	42.5	45.7	44.0

Table 8. In-domain and zero-shot retrieval performance on BEIR benchmark (Thakur et al., 2021), as measured by **Recall@100**. All the baseline number in the table are taken from (Thakur et al., 2021). ‡ indicates in-domain retrieval performance. ∗ indicates capped recall following original benchmark setup. We exclude (in-domain) MSMARCO from average computation as per common practice.

	Lexical	Sparse			Dense					
Model and size (→)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**GenQ**	**SentenceBERT** *(our teacher)*	**EmbedDistill** *(ours)*
Dataset (↓)				12-layer	12-layer	12-layer	6-layer	12-layer	12-layer	*6-layer*
MS MARCO	65.8	75.2‡	79.3‡	81.9‡	55.2	85.2‡	88.4‡	88.4‡	91.7‡	90.6‡
TREC-COVID	49.8*	34.7*	40.9*	54.1*	21.2*	45.7*	38.7*	45.6*	54.1*	48.8*
NFCorpus	25.0	23.5	24.3	25.3	20.8	23.2	28.0	28.0	27.7	26.7
NQ	76.0	63.6	78.7	83.2	88.0‡	83.6	90.3	86.2	91.1	89.9
HotpotQA	74.0	73.1	65.1	70.9	59.1	57.8	72.8	67.3	69.7	68.3
FiQA-2018	53.9	48.9	44.6	59.8	34.2	58.1	59.3	61.8	62.0	60.1
ArguAna	94.2	93.2	89.3	97.2	75.1	93.7	94.2	97.8	89.2	87.8
Touché-2020	53.8	40.6	38.1	55.7	30.1	45.8	43.1	45.1	45.3	45.5
CQADupStack	60.6	54.5	52.1	63.8	40.3	57.9	62.2	65.4	63.9	61.3
Quora	97.3	95.4	89.6	98.2	47.0	98.7	98.6	98.8	98.5	98.1
DBPedia	39.8	37.2	41.1	36.5	34.9	31.9	49.9	43.1	46.0	42.6
SCIDOCS	35.6	31.4	29.7	36.0	21.9	26.9	33.5	33.2	32.5	31.5
FEVER	93.1	73.5	84.3	91.6	84.0	90.0	93.7	92.8	93.9	93.8
Climate-FEVER	43.6	23.2	22.7	42.7	39.0	44.5	53.4	45.0	49.3	47.6
SciFact	90.8	89.3	86.3	91.4	72.7	81.6	89.1	89.3	88.9	87.2
AVG (w/o MSMARCO)	63.4	55.9	56.2	64.7	47.7	60.0	64.8	64.2	65.1	63.5

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

F. Embedding analysis

F.1. DE to DE distillation

Traditional score matching-based distillation might not result in transfer of relative geometry from teacher to student. To assess this, we look at the discrepancy between the teacher and student query embeddings for all q, q' pairs: $\|\texttt{emb}_q^{\mathrm{t}} - \texttt{emb}_{q'}^{\mathrm{t}}\| - \|\texttt{emb}_q^{\mathrm{s}} - \texttt{emb}_{q'}^{\mathrm{s}}\|$. Note that the analysis is based on NQ, and we focus on the teacher and student DE models based on BERT-base and DistilBERT, respectively. As evident from Fig. 3, embedding matching loss significantly reduces this discrepancy.

F.2. USTAD to DE distillation

We qualitatively look at embeddings from a traditional CE model (“`[CLS]`-pooled”) or USTAD CE model (“Dual-pooled”) in Fig. 4. The embedding $\texttt{emb}_{q,d}^{\mathrm{t}}$ from `[CLS]`-pooled CE model does not capture semantic similarity between query and document as it is solely trained to classify whether the query-document pair is relevant or not. In contrast, the query embeddings $\texttt{emb}_{q \leftarrow (q,d)}^{\mathrm{t}}$ from our USTAD (“Dual-pooled”) model do not degenerate and its embeddings groups same query whether conditioned on positive or negative document together. Furthermore, other related queries are closer than unrelated queries. Such informative embedding space would aid distillation to a DE model via embedding matching.

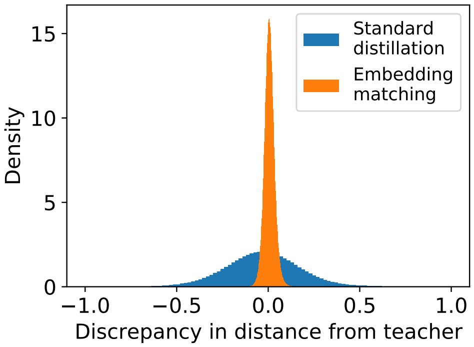


Figure 3. Histogram of teacher-student distance discrepancy in queries.

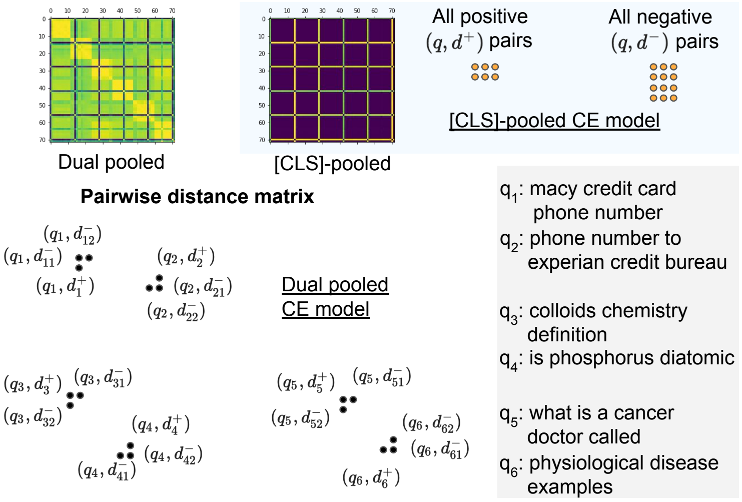


Figure 4. Illustration of geometry expressed by `[CLS]`-pooled CE and our USTAD CE (“Dual-pooled”) on 6 queries from MSMARCO and 12 passages based on pairwise distance matrix across these 72 pairs. `[CLS]`-pooled CE embeddings degenerates as all positive and negative query-document pairs almost collapse to two points and fail to capture semantic information. In contrast, our USTAD model leads to much richer representation that can express semantic information.