In Defense of Dual-Encoders for Neural Ranking

Aditya Krishna Menon [1] **Sadeep Jayasumana** [1] **Ankit Singh Rawat** [1] **Seungyeon Kim** [1] **Sashank J. Reddi** [1]
Sanjiv Kumar [1]

Abstract

Transformer-based models have proven successful in information retrieval problems, which seek to identify relevant documents for a given query. There are two broad flavours of such models: *cross-attention* (CA) models, which learn a joint query-document embedding, and *dual-encoder* (DE) models, which learn separate embeddings for the query and the document. Empirically, CA models are often more accurate, which has motivated several works that seek to bridge this performance gap. However, a fundamental question remains less explored: does this gap reflect a limitation in DE models' *capacity*, or *training* procedure? In this paper, we study this question, with three contributions. First, we establish theoretically that with a sufficiently large encoder size, DE models can capture a broad class of scores without cross-attention. Second, we show that on real-world problems, the gap between CA and DE models may be due to the latter *overfitting* to the training set. To mitigate this, we propose a distillation strategy that focuses on preserving the *ordering* amongst documents, and confirm its efficacy on neural re-ranking benchmarks.

1. Transformer-Based Neural Ranking

Information retrieval (Mitra & Craswell, 2018) is the classic problem of identifying relevant documents for a given query. Typically, such retrieval is performed in a two-step manner (Matveeva et al., 2006): one uses an initial model to efficiently *retrieve* a candidate set of documents for a query, and then uses a second model to *re-rank* these candidates. The retrieval stage is generally implemented by a model with low inference cost, for which candidate generation is tractable; if this candidate set is reasonably small, it is feasible to use more complex models for re-ranking.

Transformer-based models such as BERT (Devlin et al., 2019) have proven successful in both the retrieval and re-ranking stages. Such *neural ranking* models have two flavours: *dual-encoder* (*DE*) modelss (Lee et al., 2019; Chang et al., 2020; Karpukhin et al., 2020) which learn separate (factorised) embeddings for the query and document; and *cross-attention* (*CA*) models (Nogueira & Cho, 2019; Dai & Callan, 2019; Yilmaz et al., 2019), which learn a joint embedding for the query and document. Only DE models are applicable for retrieval, as they admit efficient nearest neighbour search (Guo et al., 2020; Johnson et al., 2021) to identify candidate documents. Both CA and DE models are applicable for re-ranking; however, empirically, CA models perform better (Hofstätter et al., 2020a).

Does the re-ranking performance gap between CA and DE models reflect a limitation in DE models' inherent *capacity*, or in their *training* procedure? While several works have explored means of improving DE models — e.g., by changing[1] the scoring layer (Khattab & Zaharia, 2020; MacAvaney et al., 2020; Hofstätter et al., 2020b), and by *distilling* predictions from a CA model (Lu et al., 2020; Izacard & Grave, 2020; Hofstätter et al., 2020a) — the root cause of the gap between CA and DE models remains elusive. This is not purely of conceptual interest: enabling usage of DE models for re-ranking is desirable, as they afford more efficient inference owing to their ability to pre-compute document embeddings (Khattab & Zaharia, 2020).

In this paper, we study this question with the aim of shedding light on the fundamental differences between CA and DE models. We further give a simple yet effective distillation strategy to improve the latter. Our contributions are:

(i) we establish theoretically that with a sufficiently large embedding dimension (and mild assumptions), any continuous ground-truth scores can be modelled by generic DE models (Proposition 3.1), and in particular by sufficiently deep transformer-based DE models (Proposition 3.2). Thus, in principle, there is no fundamental restriction in using DE versus CA models for re-ranking.

[1] Google Research, New York, USA. Correspondence to: Aditya Krishna Menon <adityakmenon@google.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] Unless otherwise stated, we use "DE" to mean a dual-encoder with dot-product scoring per (2).

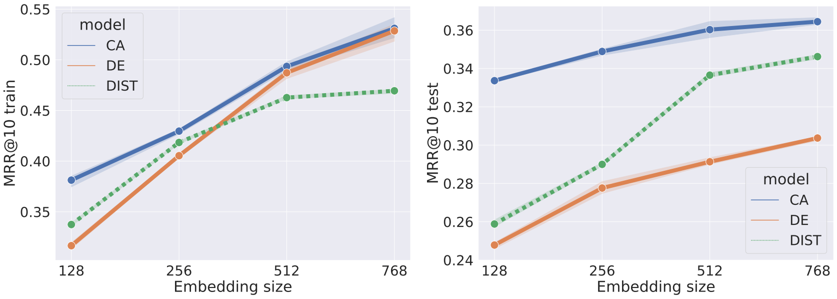

Figure 1. Comparison of cross-attention (CA) and dot-product based dual-encoder (DE) models on the MSMARCO-Passage re-ranking task. Using 6-layer BERT models with varying embedding size (Turc et al., 2019), we report the train and dev set MRR@10 averaged over three independent trials. For sufficiently large embedding dimension, the DE model closely matches the performance of the CA on the *training* set; however, there is a sizable gap in the *test* set performance. This points to the poorer DE model performance being largely an issue of generalisation, rather than model capacity. Suitable distillation (DIST) from the CA model manages to prevent such overfitting — potentially by *worsening* the *training* performance — and largely bridges the gap between the two models. See §3.2 for details on the experimental setup, and §4 for details on our distillation strategy.

(ii) we show empirically (Figure 1) that on real-world problems, CA and DE models can achieve similar *training* performance, but DE models may do worse on *test* data due to *overfitting* (§3.2). Thus, DE models may suffer due to poorer *generalisation* ability rather than *capacity*.

(iii) to mitigate the above, we propose a distillation strategy focussing on mimicking the teacher's *ordering* amongst documents (§4). This includes a generalisation of the recent margin MSE loss (Hofstätter et al., 2020a) (§4.2), and justifies the utility of softmax cross-entropy based distillation for re-ranking (Proposition 4.1, §4.3). We confirm the efficacy of this strategy on the re-reranking benchmarks MSMARCO-Passage (Nguyen et al., 2016) and Natural Questions (Kwiatkowski et al., 2019).

The above seeks to give conceptual insight into the nature of DE versus CA models for re-ranking, and suggests there is merit in further seeking ways to reducing the overfitting of DE models. In particular, the empirical Figure 1, and the theoretical Proposition 3.1, provide evidence against the hypothesis that DE models underperform on re-ranking because they are fundamentally restricted in capacity compared to CA models. Section 4 presents one means of improving DE models via distillation, offering a new M^{3}SE loss (4) that practitioners may add to their toolbox.

2. Background and Notation

We begin by formalising the retrieval problem, and the two flavours of transformer-based models for the same.

2.1. Query to Document Retrieval

Suppose we have query space $\mathcal{Q}$ and document space $\mathcal{D}$. For a given query $q \in \mathcal{Q}$, the goal of information retrieval is to identify a set of relevant documents $\mathcal{D}_{\rm rel}(q) \subseteq \mathcal{D}$ (Mitra & Craswell, 2018). This is typically achieved by learning a *scorer* $s \colon \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ that predicts the *relevance* of a query and document, and simply selecting the top-k highest scoring documents for a given query. Typically, this scorer is itself implemented in two phases (Matveeva et al., 2006; Nogueira & Cho, 2019; Chang et al., 2020). In the *retrieval* phase, one identifies an initial candidate set of documents via an initial scorer $s_{\rm ret}$. In the *re-ranking* phase, one re-scores *only* these candidate documents via a distinct scorer $s_{\rm rrk}$ to obtain $\mathcal{D}_{\rm rel}(q)$. We may thus regard $s \colon \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ as a composition of both $s_{\rm ret}$ and $s_{\rm rrk}$.

To learn a scorer, fix some parameterised class $\mathcal{S} \doteq \{s(\cdot, \cdot; \theta) \colon \mathcal{Q} \times \mathcal{D} \to \mathbb{R} \mid \theta \in \Theta\}$. Suppose we have supervision $\{(q_n, \mathbf{d}_n, \mathbf{y}_n)\}_{n=1}^N$, where for each query q_n we have a list of K associated documents $\mathbf{d}_n \doteq (d_{nk})_{k \in [K]} \in \mathcal{D}^K$ with ground-truth relevance labels $\mathbf{y}_n \doteq (y_{nk})_{k \in [K]} \in \{0, 1\}^K$. Typically, $K \ll |\mathcal{D}|$ since it is rarely feasible to collect relevance labels for *all* documents. For any given $\theta \in \Theta$, let $\mathbf{s}_n(\theta) \doteq (s(q_n, d_{nk}; \theta))_{k \in [K]} \in \mathbb{R}^K$ denote the vector of model scores on the provided documents. One may then learn a scorer s via minimising

$$R(\theta) \doteq \frac{1}{N} \sum_{n \in [N]} \ell(\mathbf{y}_n, \mathbf{s}_n(\theta))$$

for *loss* $\ell \colon \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}_+$, such as the mean square error

$$\ell_{\rm mse}(\mathbf{y}_n, \mathbf{s}_n(\theta)) \doteq \sum_{k \in [K]} (y_{nk} - s_{nk}(\theta))^2,$$

and the softmax cross-entropy (assuming for simplicity $\mathbf{1}^\top \mathbf{y} = 1$): for temperature $\tau > 0$,

$$\ell_{\rm ce}(\mathbf{y}_n, \mathbf{s}_n(\theta)) \doteq -\sum_{k \in [K]} y_{nk} \cdot \log p_{nk}(\theta) \quad (1)$$

$$p_{nk}(\theta) \doteq \frac{\exp(\tau^{-1} \cdot s_{nk}(\theta))}{\sum_{k' \in [K]} \exp(\tau^{-1} \cdot s_{nk'}(\theta))}.$$

To instantiate either objective, one missing ingredient is the precise parameterisation of $\mathcal{S}$. We now describe one possible parameterisation, given by a transformer model.

2.2. Cross-Attention and Dual-Encoder Transformers

Transformers (Vaswani et al., 2017) are sequence-to-sequence models with good empirical performance on language tasks. Suppose we have a sequence $(x_1, \ldots, x_L)$ of L *tokens* (e.g., words in a language), with $\mathbf{X} \doteq (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L) \in \mathbb{R}^{D \times L}$ being a corresponding sequence of token embeddings (e.g., word embeddings). A *transformer encoder* $T \colon \mathbb{R}^{D \times L} \to \mathbb{R}^{D \times L}$ maps this to another sequence of embeddings via a composition of *attention* and *feedforward* layers, with BERT (Devlin et al., 2019) being a canonical instantiation.

Transformer encoders can readily score relevance of a pair $(q, d) \in \mathcal{Q} \times \mathcal{D}$ of tokenised queries and documents: one may pass their concatenation through a transformer encoder T. The resulting embedding may be pooled (e.g., with a `[CLS]` token (Devlin et al., 2019)) to yield

$$s(q, d) = \mathbf{w}^\top \mathtt{pool}(T(\mathtt{concat}(q, d)))$$

for weights $\mathbf{w} \in \mathbb{R}^D$. Such *cross-attention* (CA) models apply attention layers on the queries and documents jointly, which intuitively allow for rich interactions.

A distinct strategy is to separately embed queries and documents (e.g., via BERT encoders), and score them with their dot-product (Lee et al., 2019; Chang et al., 2020; Karpukhin et al., 2020; Zhan et al., 2020; Ma et al., 2021; Xiong et al., 2021; Luan et al., 2021; Qu et al., 2021; Zhan et al., 2021a):

$$s(q, d) = \mathtt{pool}(T(q))^\top \mathtt{pool}(T(d)). \tag{2}$$

Such *dual-encoder* (DE) models have received recent interest in a range of language problems (Reimers & Gurevych, 2019; Gillick et al., 2019; Karpukhin et al., 2020). In part, this is owing to the fact that DE models can have significantly lower inference cost compared to CA models (owing to the ability to index document embeddings), which makes them appealing for both the retrieval and re-ranking stages. Indeed, DE models are widely used for retrieval, owing to their amenability to fast approximate nearest neighbour search (Guo et al., 2020; Johnson et al., 2021). Unfortunately, for re-ranking, one typically observes a large gap between DE and CA performance (see §3.2).

2.3. Knowledge Distillation

One strategy to bridge the performance gap between CA and DE models is *distillation* (Ba & Caruana, 2014; Hinton et al., 2015). Suppose we have supervision $\{(q_n, \mathbf{d}_n, \mathbf{t}_n, \mathbf{y}_n)\}_{n=1}^N$, where $\mathbf{t}_n \in \mathbb{R}^K$ are the scores from a "teacher" (e.g., CA) model. Then, we may minimise $R_{\mathrm{dist}}(\theta) \doteq \frac{1}{N} \sum_{n \in [N]} \ell(\mathbf{t}_n, \mathbf{s}_n(\theta))$, noting that we use the teacher supervision $\mathbf{t}_n$ rather than "one-hot" labels $\mathbf{y}_n$; one may also combine the teacher and one-hot labels.

Two canonical instantiations of distillation are *logit matching* (Ba & Caruana, 2014), which uses the the mean square error $\ell_{\mathrm{mse}}(\mathbf{t}_n, \mathbf{s}_n(\theta))$ between the teacher and student scores, and *probability matching* (Hinton et al., 2015), which uses the softmax cross-entropy between their softmax probabilities, i.e., $\ell(\mathbf{t}_n, \mathbf{s}_n) = \ell_{\mathrm{ce}}(\mathbf{p}_n^{\mathrm{teach}}, \mathbf{s}_n)$ where

$$p_{nk}^{\mathrm{teach}} \doteq \frac{\exp(\tau^{-1} \cdot t_{nk})}{\sum_{k' \in [K]} \exp(\tau^{-1} \cdot t_{nk'})}$$

for temperature $\tau > 0$. Empirically, distillation at least partially bridges the chasm between CA and DE performance (Lu et al., 2020; Izacard & Grave, 2020; Yang & Seo, 2020; Hofstätter et al., 2020a; Miech et al., 2021).

3. The Capacity of Dual-Encoder Models

We show that while DE models have sufficient capacity to represent complex score functions, in practice they may overfit and thus perform poorly on test data.

3.1. How Expressive are Dual-Encoders in Theory?

Our core aim is to understand the reasons behind the reranking performance gap between DE and CA models. A natural starting point concerns the *capacity* difference of these models: are DE models fundamentally unable of capturing certain query-document relevance scores?

To get a handle on this question, we begin by asking: how well can DE models approximate a given score function $s^* \colon \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$? For our purposes, the DE family comprises any model of the form $s(q, d) = \mathbf{z}(q)^\top \mathbf{w}(d)$, where $\mathbf{z} \colon \mathcal{Q} \to \mathbb{R}^D$ and $\mathbf{w} \colon \mathcal{D} \to \mathbb{R}^D$ are (arbitrarily powerful) query- and document embedding functions. Under mild assumptions on $\mathcal{D}$, we may show that DE models with sufficiently large embeddings can model a broad class of s^*.

Proposition 3.1. *Suppose $\mathcal{D}$ is a compact metric space, and $s^*(q, \cdot) \colon \mathcal{D} \to \mathbb{R}$ is continuous $\forall q \in \mathcal{Q}$. Then, $\forall \epsilon > 0$, $\exists$ $\mathbf{z}_q, \mathbf{w}_d$ of at most countably infinite dimension such that*

$$(\forall q \in \mathcal{Q}, d \in \mathcal{D}) \, |s^*(q, d) - \mathbf{z}_q^\top \mathbf{w}_d| \le \epsilon.$$

The proof (Appendix A) employs Mercer's theorem for kernel methods. Note that by symmetry, one can swap the roles of $\mathcal{Q}$ and $\mathcal{D}$ above. Further, the assumption that $\mathcal{D}$ is a compact metric space is mild: one may, e.g., identify each $d \in \mathcal{D}$ with a normalised Euclidean embedding.

We make a few remarks here. First, Proposition 3.1 emphatically does *not* intend to suggest a practical algorithm, as $\mathbf{z}_q, \mathbf{w}_q$ may be infinite-dimensional. Rather, its aim is to establish the theoretical capacity of DE models, which will prove useful in explaining certain empirical phenomena (cf. Figure 1). We shall return to this issue in §3.2.

Further, the above does not restrict the function class used to represent $\mathbf{z}$ or $\mathbf{w}$. Restricting these to transformers, we may appeal to their universal approximation power (Yun et al., 2020): a sufficiently deep transformer encoder can approximate any continuous $T\colon \mathbb{R}^{D \times L} \to \mathbb{R}^{D \times L}$ to arbitrary precision. Now suppose we represent queries $\mathcal{Q} \subset \mathbb{R}^{D \times L}$ as embeddings of sequences of L tokenised elements, with a fixed `[CLS]` token at the first position. Then, any continuous $\mathbf{z}\colon \mathcal{Q} \to \mathbb{R}^{D}$ can be approximated by reading the first token embedding of a suitable transformer T.

Proposition 3.2. *Fix some compact $\mathcal{Q} \subset \mathbb{R}^{D \times L}$. For any continuous $\mathbf{z}\colon \mathcal{Q} \to \mathbb{R}^{D}$ and $\epsilon' > 0$, $\exists$ a transformer encoder $T\colon \mathbb{R}^{D \times L} \to \mathbb{R}^{D \times L}$ with $\int_{\mathcal{Q}} \|\mathbf{z}(\mathbf{Q}) - T(\mathbf{Q})_{:1}\|_2^2 \, \mathrm{d}\mathbf{Q} \leq \epsilon'$.*

The above suggests that DE models have "high capacity", in the sense of there existing a suitable model configuration that can mimic an arbitrary (query, document) score relationship. This is *not* simply a consequence of such models having many parameters: it is also that the way these parameters are *used* — i.e., the factorised score in Equation 2 — is sufficiently powerful. To further illustrate this point, consider a model that produces scores via $w_1^\top \mathtt{pool}(T(q)) + w_2^\top \mathtt{pool}(T(d))$ for learned weights w_1, w_2. Such a model can have a large number of parameters, but will be unable to model scores that involve any interaction between the queries and documents.

Remark. Luan et al. (2021) similarly analysed the expressive power of DE models, but through a slightly different lens: they formalised the expressiveness of DE score functions based on *random projections*. This elegant analysis sheds light on one possible failure mode of DE models, i.e., scoring the relevance of queries and overly long documents. By contrast, we consider the representation power of transformer-based DE models; further, we now identify a distinct reason for the gap between DE and CA performance.

3.2. How Expressive are Dual-Encoders in Practice?

The previous section suggests that DE models ought to be sufficiently expressive as to model a broad class of scores. However, this is at odds with the wealth of empirical results suggesting a non-trivial performance gap between DE and CA models (Lu et al., 2020; Izacard & Grave, 2020; Hofstätter et al., 2020a; Miech et al., 2021). To reconcile this, note that it is a distinct question as to whether it is feasible to *learn* a good DE model from a finite number of samples, and with a capacity restriction on the function class (e.g., a fixed embedding dimension and transformer depth).

To study this, we perform an experiment on the benchmark MSMARCO-Passage dataset (Nguyen et al., 2016), which concerns query to passage retrieval; we defer a detailed discussion of this dataset and our training protocol to §5. We focus on the re-ranking setting, wherein both the CA and DE models operate on the outputs of a BM25 retrieval model (Robertson & Zaragoza, 2009). We train a series of BERT-based CA and DE models on the ("small") triplets training set, employing 6-layer BERT models from Turc et al. (2019) with varying embedding size. For each model, we compute the *mean reciprocal rank* (MRR)@10 (Radev et al., 2002) on the provided train and dev set. (We shall refer to the "dev" set as the "test" set for simplicity.)

Figure 1 compares the train and test MRR@10 for cross-attention (CA) and dual-encoder (DE) models; see Table 5 (Appendix) for a numeric summary. With increased embedding dimension, the DE model closely matches the performance of the CA on the *training* set. This aligns with §3.1, which suggests deep transformer-based DE models ought to model any ground truth scores. However, there is a sizable gap on the *test* set. This points to the DE models suffering from poorer *generalisation*, rather than *capacity*. See Appendix C.1 for results with 2-layer BERT models.

3.3. What Causes the Generalisation Gap?

We now attempt to further understand the cause for the DE model's generalisation gap. We perform a more fine-grained inspection of the CA versus DE model predictions by studying the predicted scores for positive and negative pairs. More precisely, for each query q in the dev set, we consider the distribution of scores for positive and negative (query, document) pairs (q, d^+) and (q, d^-). (Similar trends hold on the training set; see Figure 6 in the Appendix). For visual clarity, we apply a constant offset to each model's scores to ensure they are mean zero; this accounts for the *baseline shift* of DE over CA model scores (cf. Figure 7).

Figure 2 shows that both CA and DE models possess clear modes for the positive and negative pairs. However, the CA model does not "waste" its capacity by unnecessarily modelling fine-grained distinctions amongst negative pairs; rather, it collapses most negatives to a small range of scores, and focusses instead on clearly separating the positives and negatives. By contrast, the DE model has a greater overlap in the positive and negative scores, implying the model has more difficulty making distinctions between these pairs. From the right plot, the DE model has smaller *normalised margins* $(s(q, d^+) - s(q, d^-))/\rho$ between the positive and negative pairs, where ρ is the maximal score range.

The DE score distribution for negatives may be understood as follows: when updating a DE model for a given (q, d) pair, the factorised nature of the model implies a non-trivial influence on the scores for *all* other pairs (q, d') and (q', d) sharing either the query or document. Consequently, it is

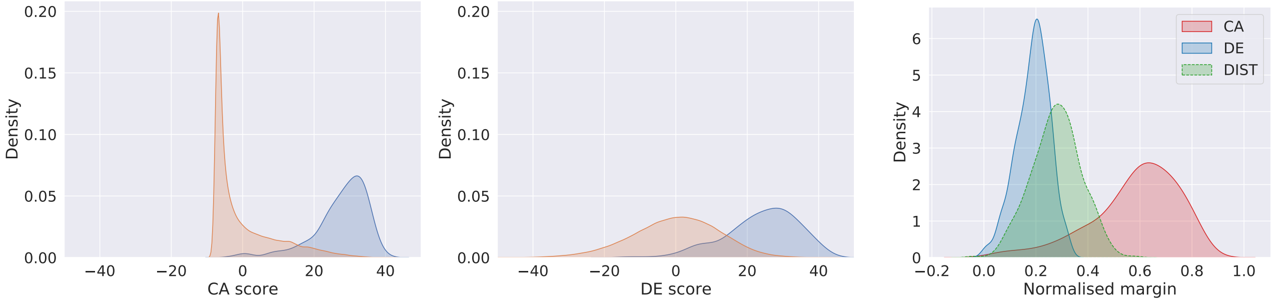

Figure 2. Comparison of CA (left) and DE (middle) model predictions on the MSMARCO-passage dev set. The CA model more confidently distinguishes positives from negatives, with a pronounced peak in scores for the latter. For visual clarity, the scores are translated to have mean zero; see Figure 7 for an uncentered plot. We further see that the *normalised margins* of the DE model are distinctly smaller than CA (right); this may be mitigated with suitable distillation (DIST).

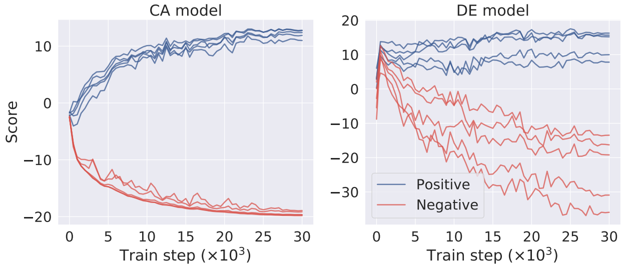

Figure 3. Evolution of scores for a sample of (query, doc) pairs under the CA and DE models across training steps. For a fixed query, we consider its 5 associated positive documents, and 5 negative documents having high token overlap with the positives. Across training steps, the CA model smoothly decreases (increases) the negative (positive) scores. However, the DE score evolution is far noisier, particularly for the negative documents. Intuitively, due to the factorised nature of the model, each update on a pair (q, d) non-trivially influences the scores for related pairs (q, d').

harder for the model to make fine-grained distinctions. This can be verified qualitatively; the DE model often makes errors on pairs with only superficial relevance (Appendix C.4).

Quantitatively, we illustrate this point as follows. We pick a single query q from the MSMARCO training set, and a set of 10 documents comprising 5 positive documents associated with q, and 5 negative documents which have high token overlap with the positives. In Figure 3, we plot the evolution of scores across training steps for each (q, d) pair under the CA and DE models. While the CA model smoothly decreases (increase) the negative (positive) scores, the DE models' updates are far noisier, particularly for the negative documents. Per the intuition above, updating based on a positive pair (q, d^+) can inadvertently increase the score for a negative pair (q, d^-), particularly when d^- has superficial similarity to d^+. The DE model thus has to repeatedly counteract these effects, resulting in a diffuse final score distribution. (See Appendix C.5 for additional plots.)

3.4. How Can We Mitigate the Generalisation Gap?

Having identified the overfitting problem plaguing DE models, one may naturally ask how to mitigate this. One natural option is to apply standard regularisation strategies, such as dropout; however, our experiments reveal these to be largely ineffective in mitigating DE overfitting (see Appendix 5.2).

The literature has identified two promising alternatives. The first is to modify the *scoring function* used to compute the DE model score based on the query and document embeddings; i.e., replace (2) with $s(q, d) = \texttt{score}(T(q), T(d))$ for suitable score (e.g., MacAvaney et al., 2020; Khattab & Zaharia, 2020; Luan et al., 2021; Santhanam et al., 2021). Such strategies have proven effective, but introduce overhead over dot-product scoring. The second strategy is to *distill* the predictions of the CA model into the DE model. By itself, this does not require changing the dot-product scoring in (2). We now study this strategy more closely.

4. Improving Dual-Encoders via Distillation

We now study CA to DE distillation schemes for re-ranking.

4.1. (Why) Does Distillation Mitigate DE Overfitting?

Several works have explored distillation from a CA "teacher" to a DE "student" (Hofstätter et al., 2020a; Yang & Seo, 2020; Miech et al., 2021), and convincingly demonstrated gains over a DE model trained on one-hot labels. The success of distillation is intuitive: smoothing the targets based on an accurate teacher ought to improve the DE models' scores compared to one-hot training.

Recall, however, that we saw in §3 that DE models' underperformance for re-ranking can be a manifestation of a *generalisation* gap (Figure 1), which in turn is a consequence of DE models having poorer separation of positive and negative pairs (Figure 2), and making noisier updates (Figure 3). These issues are not directly considered in the above works; conceptually, it is thus not clear whether distillation specifically addresses any of these limitations.

We now propose a new loss to more directly address this, and discuss how it relates to the above proposals.

4.2. Improving DE Margins via the Multi-Margin MSE

To design a suitable distillation loss for DE models, we will focus on the *margin* issue identified in Figure 2: standard DE models can produce overlapping distributions for positive and negative pairs. Intuitively, then, one simple strategy to address this is to encourage the DE student to match the CA teacher's (large) margin between positives and negatives.

To proceed, suppose each sample is the form $(q, \mathbf{d}, \mathbf{y}, \mathbf{t})$, where for query q and documents $\mathbf{d} \in \mathcal{D}^K$, we have ground-truth relevance labels $\mathbf{y} \in \{0, 1\}^K$ and "teacher" scores $\mathbf{t} \in \mathbb{R}^K$ (e.g., the outputs of a CA model). Let P, N denote the set of positive and negative documents, with $P \cup N = [K]$. To fit scores $\mathbf{s} \in \mathbb{R}^K$ for a "student" model (e.g., a DE model), one must define a suitable loss $\ell(\mathbf{t}, \mathbf{s})$. As a first attempt to encode the above intuition, consider

$$\ell_{\mathrm{mmse}}(\mathbf{t}, \mathbf{s}) \doteq \sum_{i \in P} \sum_{j \in N} ((t_i - t_j) - (s_i - s_j))^2. \quad (3)$$

While intuitive, (3) demands matching *all* documents' margins *exactly*. This may be an ineffective use of model capacity: for re-ranking, it suffices to ensure a *separation* between the positive and negative documents, thus overcoming the overlapping scores in Figure 2. To rectify this, suppose that the negative documents are sorted in descending order of the teacher scores $\mathbf{t} \in \mathbb{R}^K$. Consider now:

$$\ell_{\mathrm{m3se}}(\mathbf{t}, \mathbf{s}) = \sum_{i \in P} ((t_i - t_{j^*}) - (s_i - s_{j^*}))^2 + \sum_{j \in N} [s_j - s_{j^*}]_+^2, \quad (4)$$

where $j^* \in N$ is the index of the negative document with highest teacher score. This *multi-margin MSE* (*M*^{3}*SE*) loss only matches the margins for the highest scoring negative j^*; for all other negatives $j \neq j^*$, we simply encourage them to have lower score than j^*. We shall subsequently confirm (§5) that this loss results in a better margin distribution, and consequently achieves good re-ranking performance.

4.3. Relating the M^{3}SE to Existing Losses

We now relate Equation (4) to existing distillation losses.

Margin MSE. Suppose $K = 2$, so that each query has a *single* associated positive and negative document, with teacher and student scores $\mathbf{t} = (t_1, t_2)$, $\mathbf{s} = (s_1, s_2)$. Then, (4) reduces to the *margin MSE* (Hofstätter et al., 2020a):

$$\ell_{\mathrm{mmse}}(\mathbf{t}, \mathbf{s}) \doteq ((t_1 - t_2) - (s_1 - s_2))^2. \quad (5)$$

Equation (5) is similar to the *logit matching* or *mean square error* loss $\ell_{\mathrm{mse}}(\mathbf{t}, \mathbf{s}) = \|\mathbf{t} - \mathbf{s}\|_2^2 = (t_1 - s_1)^2 + (t_2 - s_2)^2$, but with a key difference: rather than match raw scores, one matches *margins*. This is important in settings where the range of scores for teacher and student are not commensurate, as we empirially observe with the *baseline shift* between CA and DE model scores, per §3.2 and Figure 7 (Appendix).

The M^{3}SE loss can be seen as an extension of (5) to the case of $K > 2$. This setting is intuitively desirable for re-ranking: by leveraging the teacher scores for *multiple* documents, we can more directly control their scores under the DE student.

Softmax cross-entropy. The connection to the margin MSE provides a useful foothold to relate the M^{3}SE loss to the softmax cross-entropy $\ell_{\mathrm{ce}}(\mathbf{t}, \mathbf{s})$ (1). This classical loss has been employed to improve DE models in, e.g., Lu et al. (2020); Yang & Seo (2020); Miech et al. (2021). In fact, we now show that the softmax cross-entropy is a smooth approximation to the margin MSE (5) in a precise sense.

Proposition 4.1. *Fix any teacher and student scores* $\mathbf{t}, \mathbf{s} \in \mathbb{R}^2$. *Let* $\ell_{\mathrm{sce}}(\mathbf{t}, \mathbf{s}; \tau)$ *denote the softmax cross-entropy loss* (1) *with temperature* $\tau > 0$, *and* ℓ_{mmse} *the margin MSE loss* (5). *Then, for temperature scaled sigmoid function* $\sigma(z; \tau) \doteq (1 + \exp(-\tau^{-1} \cdot z))^{-1}$, *and any* $i \in \{1, 2\}$,

$$\ell_{\mathrm{sce}}(\mathbf{t}, \mathbf{s}; \tau) = \mathrm{KL}_{\mathrm{bin}}(\sigma(t_1 - t_2; \tau) \parallel \sigma(s_1 - s_2; \tau))$$

$$\lim_{\tau \to +\infty} \left[\tau^2 \cdot \frac{\partial}{\partial s_i} \ell_{\mathrm{sce}}(\mathbf{t}, \mathbf{s}; \tau) \right] = \frac{1}{8} \cdot \frac{\partial}{\partial s_i} \ell_{\mathrm{mmse}}(\mathbf{t}, \mathbf{s}).$$

We make a few remarks. First, ℓ_{sce} swaps the square loss for a binary KL divergence $\mathrm{KL}_{\mathrm{bin}}$ in (5). Second, the softmax and margin MSE loss derivatives with respect to the student logits converge in the high temperature limit. Thus, the softmax cross-entropy is a *smooth approximation* to the margin MSE in a precise sense. Intuitively, this is plausible, as both losses are invariant to translations in the scores.

Third, our argument is subtly different from Hinton et al. (2015), which establishes convergence of the softmax and *logit matching* derivatives. The latter only holds under the assumption that the teacher and student logits are centered for *each* query; this assumption typically does *not* hold for CA and DE models, per §3.2 and Figure 7 (Appendix). By contrast, Proposition 4.1 makes no such assumption.

Finally, while Proposition 4.1 assumes $K = 2$, the softmax cross-entropy is readily applicable when $K > 2$. In this setting, we may contrast its behaviour to the M^{3}SE loss (4): in the high temperature regime, the former approaches $\max_{i \neq k^*} [s_i - s_{k^*}]_+$, where k^* is the highest scoring document (either positive or negative) under the teacher. This seeks to match the teacher's highest scoring document, but unlike (4), does not explicitly mimic its margin.

RankDistil. Equation (4) can be related to the recent

RankDistil framework (Reddi et al., 2021). Here, the goal is to design distillation losses to match the teacher and student *label ranking*, particularly for the top-k elements. A canonical instantiation is *binary RankDistil* (RankDistil-B), $\ell_{\mathrm{rankB}}(\mathbf{t}, \mathbf{s}) = \sum_{i \in P} (t_i - s_i)^2 + \sum_{j \in N} [s_j - \gamma_0]_+^2$, for constant threshold $\gamma_0 \in \mathbb{R}$. The relation between this loss and (4) is analogous to that between logit matching and margin MSE: (4) matches *margins* rather than raw scores.

4.4. Discussion and Extensions

We make two further comments on our use of distillation.

Distillation for retrieval phase. Our focus thus far has been on bridging the gap between the CA and DE models for re-ranking. However, the losses presented above are equally applicable during the retrieval phase, and are *complementary* to recent proposals for improving the retrieval performance of dual-encoder models, including using suitable hard negative mining (Xiong et al., 2021; Zhan et al., 2021a; Qu et al., 2021; Ren et al., 2021), and quantisation (Zhan et al., 2021b).

Beyond distillation? We have explored distillation to improve DE model performance, but per §3.4, there are several other options include changing the scoring function (Khattab & Zaharia, 2020), and adding hard negatives (Qu et al., 2021). Conceptually, these approaches are complementary (cf. Appendix C.6). Practically, it is of interest to understand which *specific* combination of these techniques is the most performant. However, our focus is *not* in advancing the state-of-the-art, and so we defer this study for future work.

5. Experimental Results

We now demonstrate the empirical viability of the proposed distillation scheme in §4. In particular, we show that this scheme helps reduce the gap between CA and DE models by mitigating the overfitting issue identified in §3.2.

5.1. Experimental setup

Datasets. We present results on MSMARCO-Passage (Nguyen et al., 2016) and Natural Questions (NQ) (Kwiatkowski et al., 2019). For MSMARCO, we report results on the standard dev set, and the TREC DL19 test set (Craswell et al., 2020). MSMARCO comprises (query, passage) records and their relevance labels from a human rater; for each query, the passages are retrieved from a BM25 model. The canonical training set comprises *triplets* of a query, and a single positive and negative passage. NQ comprises user questions and Wikipedia passages potentially containing their answers; we use a processed version from Karpukhin et al. (2020) (cf. Appendix B).

Models. We use transformer encoders initialised with the standard pre-trained BERT model checkpoints. Following Hofstätter et al. (2020a), we use BERT-Base for the CA model, and a 6-layer BERT model (Turc et al., 2019) with embedding size 768 for all DE models. For the DE models, we tie the query and document encoder parameters. For distillation, we use the CA model as the "teacher" and the DE model as the "student". On MSMARCO, for ease of comparison, we use the "T1" CA model annotations from Hofstätter et al. (2020a), which achieves similar re-ranking performance as our independently trained CA model. See Appendix B for further training details.

Metrics. We report the MRR@10 and nDCG@10 for all methods. Note that we are primarily interested in *re-ranking* tasks, wherein the test set comprises documents retrieved by a BM25 baseline, and our models re-rank these candidates. For MSMARCO, we additionally consider full retrieval performance obtained by scoring *all* (8.8M) passages, but reiterate that maximising performance on this metric requires additional negative mining.

Baselines. We consider methods that leverage either the observed ("one-hot") training labels on the triplet data, or the predictions from the cross-attention ("teacher") model on the triplet data. For the former, we employ softmax cross-entropy against the using a cross-attention and dual-encoder model. For the latter, we employ the logit MSE loss (Ba & Caruana, 2014), margin MSE loss (Hofstätter et al., 2020a), and the binary version of RankDistil (RankDistil-B) (Reddi et al., 2021). These are compared against the proposed M^{3}SE (Equation 4), and the softmax cross-entropy (Equation 1). The latter methods handle of an arbitrary number of documents K in the supervision (i.e., they are not restricted to triplet data). We thus aggregate the triplet data by query, and retain the top-20 passages with highest teacher model score. We study the sensitivity to K in Appendix D.3.

Scope of results. Before proceeding, we emphasise the aim and scope of our results. First, per §4.4, our primary goal is to explore the feasibility of bridging the gap between CA and DE models for *re-ranking*. This is distinct from the more common use of DE models in the literature for *retrieval* (Chang et al., 2020; Xiong et al., 2021; Zhan et al., 2021a); recall that CA models are not applicable for this setting. While we do not make special effort to incorporate tricks from the retrieval literature (e.g., using hard negatives), our proposed distillation techniques can improve performance for this stage as well; see Appendix E.2.

Second, our focus is in advancing their conceptual understanding of DE models for re-ranking, rather than the state-of-the-art. Methods achieving the latter involve ideas such as careful selection of hard negatives (Ren et al., 2021), and scoring functions beyond the dot-product (Santhanam et al., 2021); each of these is potentially complementary to our

	MSMARCO re-rank		TREC DL19 re-rank		NQ re-rank		
Model	MRR	nDCG	MRR	nDCG	MRR	nDCG	
One-hot models							
BM25 (Robertson & Zaragoza, 2009)	0.194†	0.241†	0.689†	0.501†	—	—	
ANCE (Xiong et al., 2021)	—	—		—	0.677†	—	—
Cross-attention BERT (12-layer)	0.370	0.430	0.829	0.749	0.746	0.673	
Dual-encoder BERT (6-layer)	0.310	0.360	0.834	0.677	0.676	0.601	
Distilled dual-encoders							
MSE (Hofstätter et al., 2020a)	0.289	0.343	0.781	0.693	0.659	0.591	
Margin MSE (Hofstätter et al., 2020a)	0.334	0.392	0.867$^\diamond$	0.718	0.673	0.594	
RankDistil-B (Reddi et al., 2021)	0.249	0.301	0.852	0.708	0.649	0.561	
Softmax CE (Equation 1)	0.346	0.405	0.846	0.726$^\diamond$	0.682	0.607	
M^{3}SE (Equation 4)	0.349	0.406	0.852	0.714	0.699	0.625	

Table 1. Summary of MRR@10 and nDCG@10 on MSMARCO Passage and Natural Questions (NQ). We compare cross-attention, dual-encoder (DE), and distilled dual-encoder BERT models. We highlight the best performing DE-based model. Distilling the DE model with our proposed techniques consistently improves performance. Results marked † are quoted from the corresponding reference, and "—" are not available from the reference. $^\diamond$ See Appendix E.3 for analysis of potential label noise influencing the results.

use of different distillation losses (see Appendix C.6).

5.2. Results and Discussion

We discuss salient findings from the results in Table 1.

Baseline performance. Comparing the performance of CA and DE models in Table 1, we make two initial observations. First, there is a sizable performance gap between the models when using one-hot labels. Second, in line with Hofstätter et al. (2020a), naïve logit matching via the MSE underperforms, due to the vastly different scales of the two models; this is however assuaged by the margin MSE loss.

Efficacy of distillation. Both M^{3}SE and the softmax cross-entropy significantly outperform existing losses on the MSMARCO and NQ re-reranking tasks, and shrinks the gap between the DE and CA models. On the TREC DL19 test set, our proposed methods appear to fall short of the margin MSE MRR@10; a closer inspection of the apparent loss cases (Appendix E.3) reveals these are unanimously the results of false negatives in the provided labels.

Why does distillation help? Figure 1 shows that M^{3}SE distillation nearly closes the gap between CA and DE models. We now attempt to evince *why* this strategy helps. First, Figure 2 (right) shows that softmax cross-entropy distillation makes the DE scores better behaved: it shifts the margins of the DE model to be closer to the the CA model.

Second, observe that the *test* gains comes with some sacrifice of *training* MRR for the highest embedding size. This is consistent with observations in the classical distillation literature (Cho & Hariharan, 2019). Figure 5 (Appendix) further shows the learning curve on the MSMARCO train and test set, which demonstrates a consistent gap in test set performance as training progresses. Distillation yields a training MRR that closely matches that of the DE model for the first $100K$ training steps; subsequently, however, the training MRR of the distilled model saturates, while that of the one-hot model keeps increasing. Intuitively, distillation prevents fitting to noisy samples beyond this point.

Beyond dot-product DE models. We have thus far focussed on standard dot-product based scoring for the DE models, per (2). It is of interest as to what impact more complex scoring functions, per §3.4, have on final model performance. To study this, we consider the ColBERT model (Khattab & Zaharia, 2020), which computes the average of the maximum query-document token similarities.

Consistent with Khattab & Zaharia (2020), using this model with one-hot labels by itself improves performance significantly over the dot-product (see Table 2). When further combined with distillation, the performance *exceeds that of the CA teacher*, reaching an MRR@10 of $\mathbf{0.376}$ for MSMARCO re-ranking under the softmax cross-entropy loss. (Full results in Appendix, Table 7.) This further highlights that factorised representations need not imply a loss in performance, provided they are suitably trained.

Do we need a CA teacher? The results thus far have relied on predictions provided by a CA teacher; however, one may naturally ask whether a suitable DE teacher can act in its place. We thus consider distilling from two DE teachers: one using the standard dot-product, and another using the ColBERT scoring function. Table 2 shows that on MSMARCO, distilling from a ColBERT teacher suffices to significantly improve over regular one-hot training; indeed, for a dot-product DE student, we *match* the performance of

	Scoring function	
Teacher	Dot	ColBERT
One-hot	0.310	0.356
Dot	0.316	0.351
ColBERT	0.334	0.368
CA	0.334	0.376

Table 2. MRR@10 of various distilled DE models on MSMARCO Passage re-ranking. We consider different teachers, and standard dot-product and ColBERT based DE models. Remarkably, we find that even when using a ColBERT teacher, we can significantly improve performance over training on one-hot labels.

distilling from a CA teacher. This illustrates that one can partly bridge the gap between DE and CA models *without* any access to a CA model in the training pipeline.

Effect of alternate regularisation strategies. Distillation provides one means of preventing the DE model from overfitting to the training set. One may of course conceive of other plausible strategies, such as:

- increasing the dropout rate in the final layer of the transformer. Our main experiments employ 10% dropout following prior work, which improves performance considerably over having no dropout; it is thus of interest whether there are benefits to further increasing this.
- performing token dropout at the input layer. Intuitively, such corruption prevents the model from overly relying on individual tokens, thus improving generalization.
- adding a masked language model (Devlin et al., 2019) loss. Adding such a self-supervised objective to training can plausibly improve the model's robustness.
- modifying the base loss from square or log-loss to the focal loss (Lin et al., 2017), which has proven effective in vision tasks involving high class-imbalance. In our setting, there are typically only a handful of relevant documents, but several millions of irrelevant documents.

Table 3 summarises results for each of these strategies on MSMARCO. Here, we focus on their effect on the standard (one-hot) training of a DE model on triplet data. Unfortunately, *none* of these strategies significantly improve the generalization performance of the baseline model. Thus, mitigating the overfitting observed in Figure 1 requires more effort than appealing to standard classification strategies.

Effect of label smoothing. A special case of distillation is label smoothing (Szegedy et al., 2016), which can be understood as using a teacher that predicts the uniform distribution over all labels, and combining the result with the one-hot label. Specifically, this involves mixing the one-hot labels with a uniform distribution over all labels, with the mixing controlled by weight $\alpha \in [0, 1)$; $\alpha = 0$ corresponds to standard one-hot training. Label smoothing has also proven effective as a means of preventing harmful overfitting, and thus improving generalisation (Müller et al., 2019; Lukasik et al., 2020) in classification settings.

Strategy	Train MRR@10	Test MRR@10
Baseline DE	0.619	0.310
Increased embedding dropout	0.588	0.299
Token dropout	0.572	0.291
Masked language loss	0.548	0.299
Focal loss	0.546	0.307

Table 3. Results of various regularisation strategies on MSMARCO Passage dev set. For all rows, we use a DE model trained on the triplet data.

We studied the effect of varying degrees of label smoothing for DE models on MSMARCO, using the softmax CE loss on triplet data. We use a 6-layer BERT model as before; for computational ease, we report results after $300,000$ steps of training, at which point all models see stable test error. Table 4 reveals that, surprisingly, smoothing has a minimal effect on test performance, even at high smoothing levels. This is despite smoothing reducing the training performance at higher values, which is expected given its equivalence to injecting symmetric label noise to each sample.

Smoothing α	Train MRR@10	Test MRR@10
0.0	0.394	0.308
0.1	0.460	0.297
0.2	0.455	0.296
0.3	0.453	0.292
0.9	0.382	0.292
0.99	0.305	0.278

Table 4. Effect of label smoothing on DE models, on the re-ranking task for the MSMARCO-Passage dataset. For all models, we use the softmax CE loss on triplet data.

Additional results. In the Appendix, we present several additional results and ablations, including results in the retrieval as opposed to re-ranking setting (Appendix E.2).

6. Conclusion and Future Work

Given the empirical observation that DE models underperform CA models for re-reranking, we have sought to understand *why* this is the case. Our results provide evidence against the hypothesis that this is because DE models are fundamentally restricted in capacity compared to CA models, and point instead to the issue being one of generalisation. Several avenues for future work remain open, e.g., studying the efficacy of the proposed losses when combined with recent advances for retrieval (Xiong et al., 2021).

References

Ba, J. and Caruana, R. Do deep nets really need to be deep? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Chang, W., Yu, F. X., Chang, Y., Yang, Y., and Kumar, S. Pre-training tasks for embedding-based large-scale retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019.

Craswell, N., Mitra, B., Yilmaz, E., and Campos, D. Overview of the TREC 2020 deep learning track. In Voorhees, E. M. and Ellis, A. (eds.), *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

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.

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.

Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, E., and Garcia-Olano, D. Learning dense representations for entity retrieval, 2019.

Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., and Kumar, S. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3887–3896. PMLR, 2020.

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, 2020a. URL https://arxiv.org/abs/2010.02666.

Hofstätter, S., Zlabinger, M., and Hanbury, A. Interpretable & time-budget-constrained contextualization for re-ranking. In Giacomo, G. D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., and Lang, J. (eds.), *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 513–520. IOS Press, 2020b.

Izacard, G. and Grave, E. Distilling knowledge from reader to retriever for question answering, 2020.

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

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 2020. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, C. P. M. Z. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of NAACL*, 2022.

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., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

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.

Lin, T., Goyal, P., Girshick, R. B., He, K., and Dollár, P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2999–3007. IEEE Computer Society, 2017.

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*, pp. 2645–2652, New York, NY, USA, 2020.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 04 2021. ISSN 2307-387X.

Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6448–6458. PMLR, 2020.

Ma, X., Sun, K., Pradeep, R., and Lin, J. A replication study of dense passage retriever. *CoRR*, abs/2104.05740, 2021. URL https://arxiv.org/abs/2104.05740.

MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., and Frieder, O. *Efficient Document Re-Ranking for Transformers by Precomputing Term Representations*, pp. 49–58. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450380164.

Matveeva, I., Burges, C., Burkard, T., Laucius, A., and Wong, L. High accuracy retrieval with multiple nested ranker. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pp. 437–444, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933697.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006.

Miech, A., Alayrac, J., Laptev, I., Sivic, J., and Zisserman, A. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9826–9836. Computer Vision Foundation / IEEE, 2021.

Mitra, B. and Craswell, N. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018. ISSN 1554-0669. doi: 10.1561/1500000061.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 4696–4705, 2019.

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., Ábrego, G. H., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M., and Yang, Y. Large dual encoders are generalizable retrievers. *CoRR*, abs/2112.07899, 2021. URL https://arxiv.org/abs/2112.07899.

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

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.

Radev, D. R., Qi, H., Wu, H., and Fan, W. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

Reddi, S., Kumar Pasumarthi, R., Menon, A., Singh Rawat, A., Yu, F., Kim, S., Veit, A., and Kumar, S. RankDistil: Knowledge distillation for ranking. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2368–2376. PMLR, 13–15 Apr 2021.

Reimers, N. 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 https://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 *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.

Robertson, S. and Zaragoza, H. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669.

Ruiyang Ren, Yingqi Qu, J. L. W. X. Z. Q. S. H. W. H. W. and Wen, J.-R. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of EMNLP*, 2021.

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.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer New York, New York, NY, 2008. ISBN 978-0-387-77241-7.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826, 2016.

Turc, I., Chang, M., Lee, K., and Toutanova, K. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962, 2019. URL http://arxiv.org/abs/1908.08962.

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.

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.

Yang, S. and Seo, M. Is retriever merely an approximator of reader? *CoRR*, abs/2010.10999, 2020. URL https://arxiv.org/abs/2010.10999.

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.

Zhan, J., Mao, J., Liu, Y., Zhang, M., and Ma, S. RepBERT: Contextualized text embeddings for first-stage retrieval. *CoRR*, abs/2006.15498, 2020. URL https://arxiv.org/abs/2006.15498.

Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., and Ma, S. Optimizing dense retrieval model training with hard negatives. *CoRR*, abs/2104.08051, 2021a. URL https://arxiv.org/abs/2104.08051.

Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., and Ma, S. Jointly optimizing query encoder and product quantization to improve retrieval performance. In *CIKM*, 2021b.

A. Proofs

Proof of Proposition 3.1. For any $q \in \mathcal{Q}$, let us write $s_q \doteq s(q, \cdot) \colon \mathcal{D} \to \mathbb{R}$. Pick a reproducing kernel Hilbert space $\mathcal{H}$ over $\mathcal{D}$ with associated inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Suppose further $\mathcal{H}$ has associated kernel $k \colon \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ that is continuous, and universal (Micchelli et al., 2006). Then, $\mathcal{H}$ is dense in the space of continuous functions (Sriperumbudur et al., 2011). Recall that by assumption s_q is continuous. Consequently, each s_q can be approximated to arbitrary precision (with respect to the ℓ_∞ metric) by some $s^*_q \in \mathcal{H}$. Since $s^*_q \in \mathcal{H}$, by the reproducing property of an RKHS,

$$(\forall d \in \mathcal{D})\, s^*(q, d) = \langle s^*_q, k_d \rangle_{\mathcal{H}},$$

where in an abuse of notation $k_d \doteq k(d, \cdot) \in \mathcal{H}$. Observe that we may interpret s^*_q as an "embedding" for the query q, and k_d a "weight vector" for the document d.

It may appear at this stage that we are done; however, each of s^*_q and k_d may be uncountably infinite dimensional objects, and the inner product in $\mathcal{H}$ may not be the standard dot-product. Nonetheless, under the stated assumptions (in particular, compactness of $\mathcal{D}$), we may appeal to a Mercer representation of the kernel, following Steinwart & Christmann (2008, Theorem 4.51). In particular, we may construct a *Mercer feature map* $\Phi \colon \mathcal{D} \to \ell_2$ for $\mathcal{H}$, such that

$$(\forall h \in \mathcal{H})\, (\exists (a_n)_{n=1}^{\infty} \in \ell_2)\, h(d) = \sum_{n=1}^{\infty} a_n \cdot \Phi_n(d).$$

Further, the inner product between any $h_1, h_2 \in \mathcal{H}$ with "Mercer coefficients" $(a_n)_{n=1}^{\infty}, (b_n)_{n=1}^{\infty}$ is expressible as

$$\langle h_1, h_2 \rangle_{\mathcal{H}} = \sum_{n=1}^{\infty} a_n \cdot b_n,$$

i.e., a familiar dot-product. In our context, write the Mercer coefficients for s^*_q as $(z_{qn})_{n=1}^{\infty}$, and for k_d as $(w_{dn})_{n=1}^{\infty}$. Thus, we may write

$$\langle s^*_q, k_d \rangle_{\mathcal{H}} = \mathbf{z}_q^\top \mathbf{w}_d,$$

where the vectors $\mathbf{z}_q, \mathbf{w}_d$ have at most countably infinite dimension.

□

Proof of Proposition 3.2. Any continuous $\mathbf{z} \colon \mathbb{R}^{D \times L} \to \mathbb{R}^D$ can be trivially associated with $\tilde{\mathbf{z}} \colon \mathbb{R}^{D \times L} \to \mathbb{R}^{D \times L}$, where each column $\tilde{\mathbf{z}}_{:i} = \mathbf{z}$. Such a mapping preserves continuity. By Yun et al. (2020, Theorem 2), we may thus find a transformer encoder $\mathbf{T} \colon \mathbb{R}^{D \times L} \to \mathbb{R}^{D \times L}$ such that $\int_{\mathcal{Q}} \|\tilde{\mathbf{z}}(\mathbf{Q}) - \mathbf{T}(\mathbf{Q})\|_2^2 \,\mathrm{d}\mathbf{Q} \leq \epsilon'$. Consequently, we must also have $\int_{\mathcal{Q}} \|\tilde{\mathbf{z}}(\mathbf{Q})_{:1} - \mathbf{T}(\mathbf{Q})_{:1}\|_2^2 \,\mathrm{d}\mathbf{Q} \leq \epsilon'$ almost surely, which by definition implies $\int_{\mathcal{Q}} \|\mathbf{z}(\mathbf{Q}) - \mathbf{T}(\mathbf{Q})_{:1}\|_2^2 \,\mathrm{d}\mathbf{Q} \leq \epsilon'$. □

Proof of Proposition 4.1. For temperature $\tau > 0$, simple algebra reveals (1) to be

$$\begin{aligned}
\ell_{\mathrm{sce}}(\mathbf{t}, \mathbf{s}) &= -\sum_{y \in \{1,2\}} \frac{\exp(\tau^{-1} \cdot t_y)}{\exp(\tau^{-1} \cdot t_1) + \exp(\tau^{-1} \cdot t_2)} \cdot \log \frac{\exp(\tau^{-1} \cdot s_y)}{\exp(\tau^{-1} \cdot s_1) + \exp(\tau^{-1} \cdot s_2)} \\
&= \sum_{z \in \{\pm 1\}} \frac{1}{1 + e^{z \cdot \tau^{-1} \cdot (t_2 - t_1)}} \cdot \log(1 + e^{z \cdot \tau^{-1} \cdot (s_2 - s_1)}) \\
&= -\sum_{z \in \{\pm 1\}} \sigma(z \cdot (t_1 - t_2); \tau) \cdot \log(\sigma(z \cdot (s_1 - s_2); \tau)).
\end{aligned}$$

This is equivalent to $\mathrm{KL}_{\mathrm{bin}}(\sigma(t_1 - t_2; \tau) \,\|\, \sigma(s_1 - s_2; \tau))$, i.e., swapping out the square loss for the KL divergence in (5). In fact, one can make a stronger connection: observe that

$$\begin{aligned}
\frac{\partial}{\partial s_1} \ell_{\mathrm{mmse}}(\mathbf{t}, \mathbf{s}) &= 2 \cdot ((s_1 - s_2) - (t_1 - t_2)) \\
\frac{\partial}{\partial s_1} \ell_{\mathrm{sce}}(\mathbf{t}, \mathbf{s}) &= -\sum_{z \in \{\pm 1\}} \sigma(z \cdot (t_1 - t_2); \tau) \cdot \frac{\partial}{\partial s_1} \log(\sigma(z \cdot (s_1 - s_2); \tau)
\end{aligned}$$

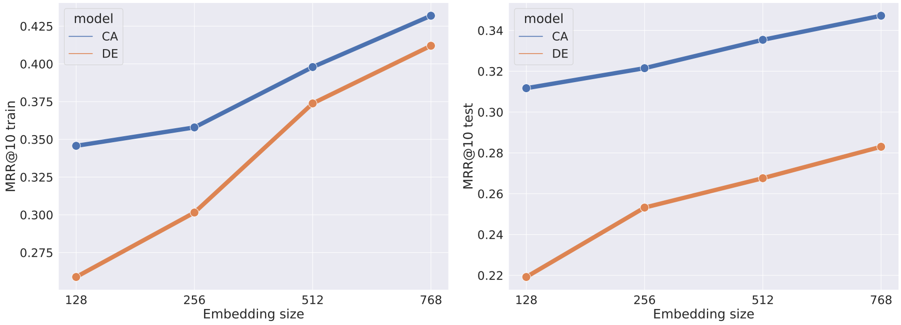

Figure 4. Comparison of BERT-based cross-attention (CA) and dual-encoder (DE) models on the MSMARCO-Passage re-ranking task. For varying embedding dimension of an underlying 2-layer BERT model, we report the train and dev set MRR@10. We see the same trend as in Figure 1: for sufficiently large embedding dimension, the DE model closely matches the performance of the CA on the *training* set; however, there is a sizable gap in the *test* set performance.

$$
\begin{aligned}
&= \tau^{-1} \cdot \left(\frac{e^{\tau^{-1} \cdot s_1}}{e^{\tau^{-1} \cdot s_1} + e^{\tau^{-1} \cdot s_2}} - \frac{e^{\tau^{-1} \cdot t_1}}{e^{\tau^{-1} \cdot t_1} + e^{\tau^{-1} \cdot t_2}} \right) \\
&= \tau^{-1} \cdot \left(\frac{1}{1 + e^{\tau^{-1} \cdot (s_2 - s_1)}} - \frac{1}{1 + e^{\tau^{-1} \cdot (t_2 - t_1)}} \right) \\
&\sim \tau^{-2} \cdot \left(\frac{s_1 - s_2}{4} - \frac{t_1 - t_2}{4} \right),
\end{aligned}
$$

as $\tau \to +\infty$, where the last line follows from the Taylor series approximation

$$
\frac{1}{1 + \exp(-z)} = \frac{1}{2} + \frac{z}{4} + \mathcal{O}(z^2)
$$

as $z \to 0$. □

B. Experiment hyperparameters

For all models, at the output layer we apply dropout at rate 0.1 and layer normalisation. We use a sequence length of 30 for queries, and 200 for passages. We optimise all methods for a maximum of 3×10^5 steps using Adam with weight decay, with a batch size of 128 and a learning rate of 2.8×10^{-5} (i.e., a $4\times$ scaling of the choices in Hofstätter et al. (2020a)). Following Hofstätter et al. (2020a), we perform early stopping based on the nDCG@10 metric.

For the Natural Questions dataset, we use the processed version used in (Karpukhin et al., 2020), that contains questions, positive passages with a correct answer to each question, and a corpus of all Wikipedia passages. To train DE, CA, and multi-negative DIST models, we use 19 BM-25 hard-negative passages for each question along with a positive passage. For single-negative DIST models we use a single BM-25 negative from the same collection. To calculate MRR@10 and nDCG@10 metrics, we use the queries in the dev set with 200 passages containing positives, 100 BM-25 hard-negatives and up to 100 random negatives.

C. Additional experiments: CA versus DE models

C.1. CA versus DE models: effect of base architecture

Figure 4 presents an an analogue of Figure 1, but using instead a 2-layer BERT model. As before, we initialise using the pre-trained models developed in Turc et al. (2019). The same general trends hold: for sufficiently large embedding dimension, the DE model closely matches the performance of the CA on the *training* set; however, there is a sizable gap in the *test* set performance. Table 5 presents a numeric summary of the train and test performance for CA and DE models, under different choices of base BERT-model architecture (in terms of number of layers, as well as the final embedding dimension).

	MRR@10 train		**MRR@10 test**	
Model	CA	DE	CA	DE
small-bert-2-128	0.356	0.292	0.309	0.221
small-bert-2-256	0.358	0.302	0.322	0.253
small-bert-2-512	0.398	0.374	0.335	0.268
small-bert-2-768	0.429	0.406	0.334	0.281
small-bert-6-128	0.385	0.317	0.334	0.249
small-bert-6-256	0.427	0.407	0.350	0.282
small-bert-6-512	0.493	0.491	0.360	0.293
small-bert-6-768	0.522	0.518	0.364	0.310
bert-base	0.658	0.635	0.368	0.309

Table 5. Comparison of cross-attention (CA) and dot-product based dual-encoder (DE) models on the MSMARCO-Passage re-ranking task. Notably, for sufficiently large embedding dimension, the DE model closely matches the performance of the CA on the *training* set; however, there is a sizable gap on the *test* set. This points to the poorer DE model performance being largely an issue of generalisation, rather than capacity.

C.2. CA versus DE models: evolution of train and test performance

The above results summarise the results at the completion of training. For a finer-grained understanding of how performance evolves during training, Figure 5 presents the learning curves for CA, DE, and distilled DE (DIST) models on the train and test set. Here, we use the small-bert-6-768 architecture for all models. We observe that the CA and DE models initially have a non-trivial gap in performance, but this shrinks over time; this further points to the initialisation of the CA models being more favourable for generalisation. Note also that beyond a certain point, both models continually improve their *training* performance, but are completely neutral in terms of *test* performance. Interestingly, distillation largely tracks the DE model performance, but then *worsens* on the training set while *improving* on the test set. This suggests that distillation helps explore a better part of the search space.

C.3. CA versus DE models: comparison of score distributions

Figure 6 compares the score distributions for the CA and DE models on the *training set*. We observe the same general trends as the test set (Figure 2). Figure 7 further presents the test score distributions without zero-centering. We see that the DE model exhibits a strong baseline shift, wherein the scores are of the order of ~ 700.

Figure 8 studies the score distributions for the CA and DE models after varying numbers of training steps. At initialisation, as expected, both models have essentially overlapping distributions for the positives and negatives. However, here too, we observe that the DE model scores operate on a much wider range than the DE model. The middle row continues this study by looking at behaviour after $10,000$ steps of training, which is a small fraction of the $300,000$ steps used for training of all models. Here, we start to observe a clearer separation between the positive and negative scores for both models. Interestingly, the CA model already sees a sharp peak on the negatives, unlike the DE model. The bottom row shows the test set score distributions of the final trained models, to complement Figure 2 from the body. We see the same general trend as the training set: the CA model is seen to more confidently distinguish positives from negatives, and operate at a narrower range of scores.

C.4. CA versus DE models: qualitative analysis

Table 6 provides a sample of (query, passage) pairs from MSMARCO-Passage dev set, where there is a large discrepancy between the CA and DE model scores. Specifically, we consider pairs that the CA model scores low, but the DE model scores high. Interestingly, these pairs typically involve strong token overlap between the query and passage — indicating a certain degree of topicality — but are fact genuine negative pairs. This reflects that the DE model may be unable to capture

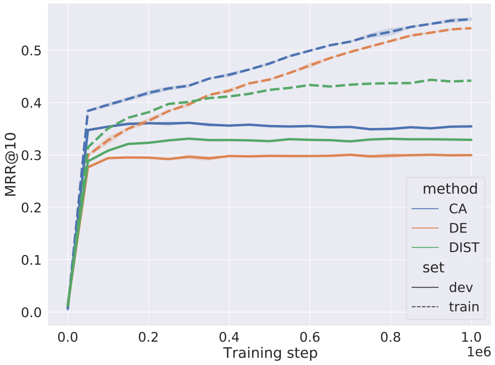

Figure 5. Learning curve for CA, DE, and DIST models on the MSMARCO-Passage train and test set. Here, we use the small-bert-6-768 architecture. Distillation is seen to saturate training performance beyond a certain point, while still resulting in a solution with better generalisation on the test set.

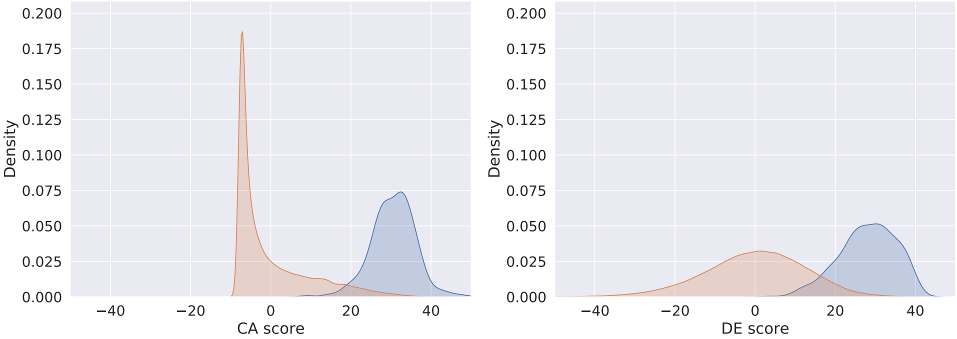

Figure 6. Comparison of CA and DE model predictions on the MSMARCO-Passage train set. We observe the same general trends as the test set (Figure 2). For visual clarity, the scores are translated to have mean zero.

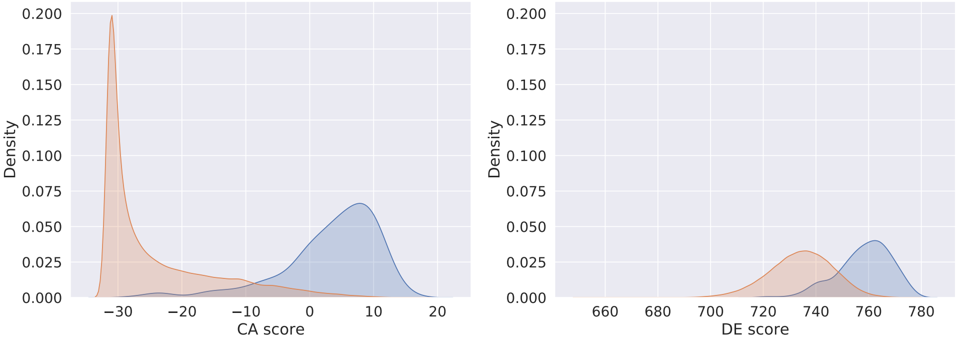

Figure 7. Comparison of CA and DE model predictions on the MSMARCO-Passage test set. Here, we do *not* center the scores to have mean zero. We see that the DE model scores possess a strong baseline shift over the model scores.

certain fine-grained distinctions.

C.5. Evolution of scores

We supplement Figure 3 with plots further illustrating the differing nature of CA and DE models. Unless otherwise specified, these plots use the query “`causes and treatment of whiteheads on face`”.

Low token-overlap documents. Figure 9 presents results when we consider 5 negative documents with the *lowest* average

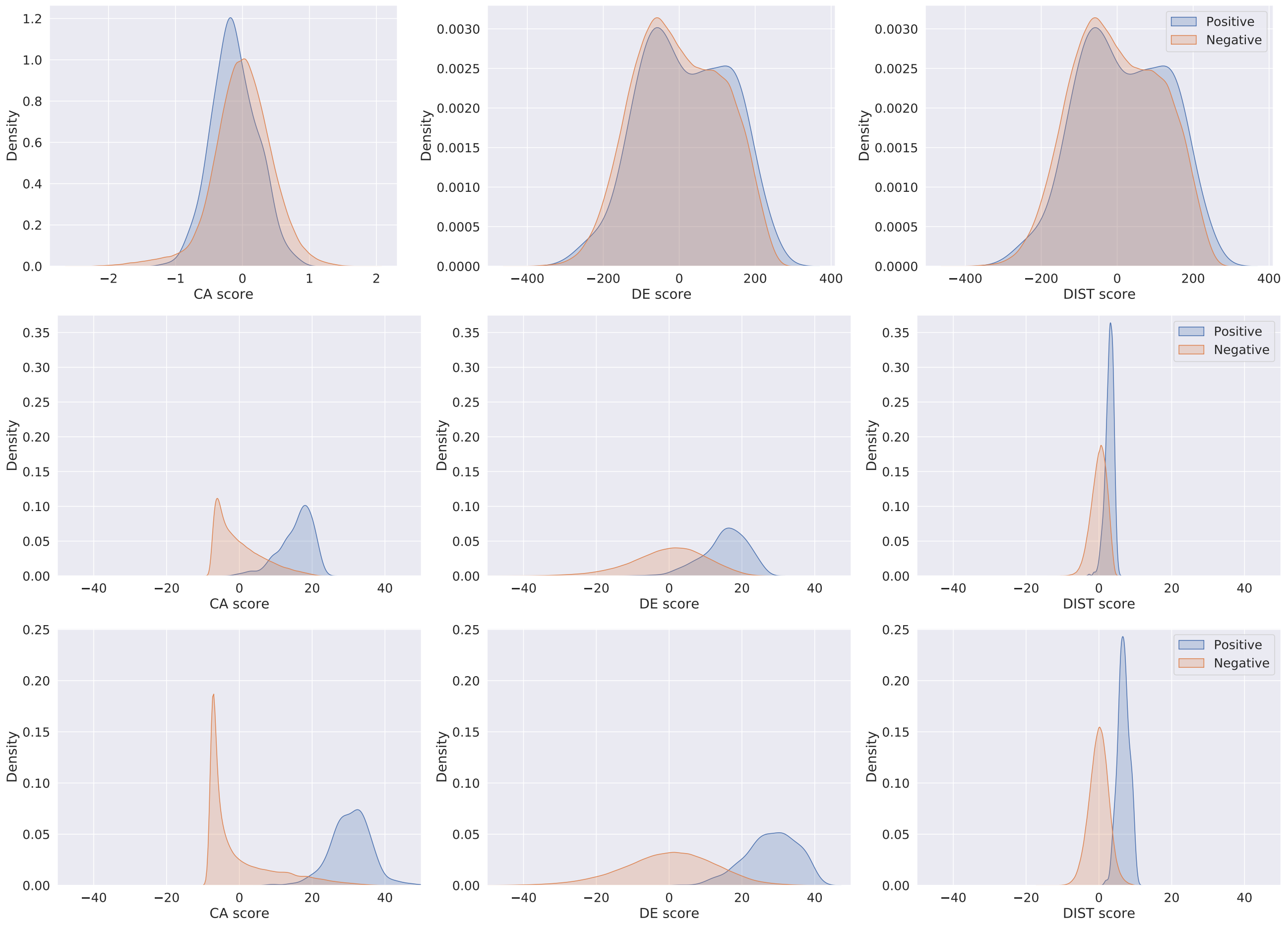

Figure 8. Comparison of CA, DE and distilled DE (DIST) model scores on the MSMARCO-Passage dev set at initialisation (top row), after $10,000$ steps of training (middle row), and at the completion of $300,000$ steps of training (bottom row). At initialisation, all models have overlapping score distributions, but the CA model operates at a much tighter range. After a few steps of training, the CA model already sees a peaky score distribution on the negative, while the DE model has a diffuse distribution. The distilled model however manages to overcome this, and produce a sharper distribution than even CA.

token overlap to the positives. Observe here that both model updates for the negatives are even smoother than the high-overlap case. However, the DE model is still considerably noisier compared to the CA model, again reflecting its difficulty in isolating score updates to a single pair.

Negative documents. Figure 10 presents results when we consider 5 negative documents which are randomly sampled from the set of top-1000 BM25 retrieved documents. Intuitively, these are harder than random documents, and some of these may indeed be false negatives. As a result, we observe that even the CA model has far noisier updates to its scores. Nonetheless, we again observe that the DE model scores evolve more inconsistently.

Different choice of query. The results thus far have considered the same query. Figure 11 illustrates results for 5 other queries, with largely the same trends as above.

False negatives. Figure 12 illustrates results for a different choice of query (“`foot pain common causes`”.), wherein the 5 negative documents are in fact largely false negatives. We now see an exacerbation of the trend in Figure 10, with the CA model having difficulty in smoothly reducing the scores.

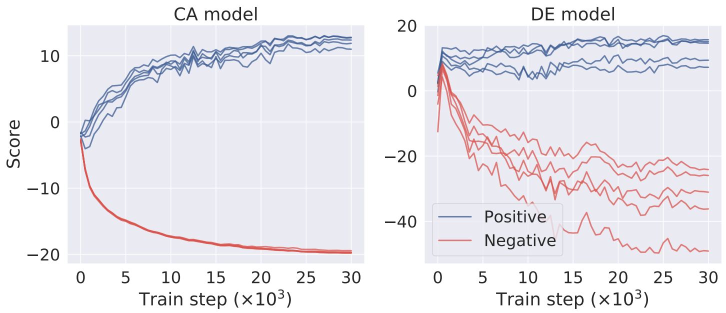

Figure 9. Evolution of scores for a sample of (query, doc) pairs under the CA and DE models across training steps. For a fixed query, we consider its 5 associated positive documents, and 5 negative documents having *low* token overlap with the positives.

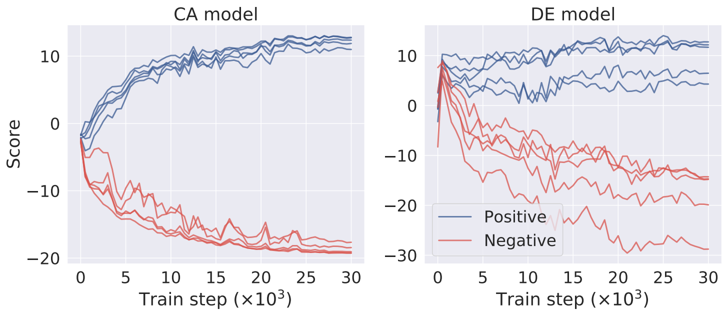

Figure 10. Evolution of scores for a sample of (query, doc) pairs under the CA and DE models across training steps. For a fixed query, we consider its 5 associated positive documents, and 5 negative documents randomly sampled from the set of top-1000 BM25 documents.

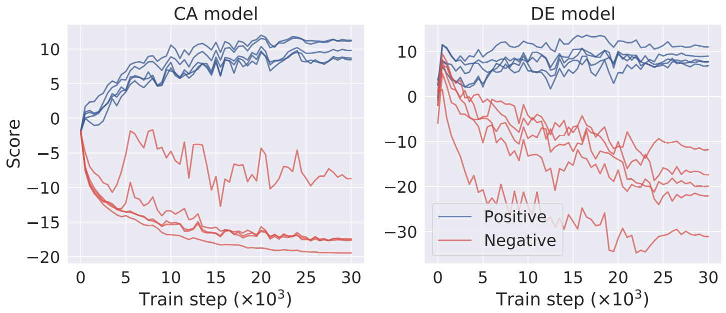

Figure 11. Evolution of scores for a sample of (query, doc) pairs under the CA and DE models across training steps. For each query from a set of 5 queries, we consider its 5 associated positive documents, and 5 negative documents sampled from the top-1000 BM25 documents.

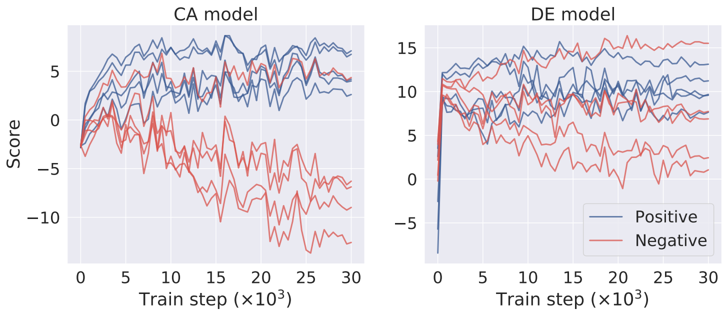

Figure 12. Evolution of scores for a sample of (query, doc) pairs under the CA and DE models across training steps. For a fixed query, we consider its 5 associated positive documents, and 5 negative documents sampled from the set of top-1000 BM25 documents. Here, the 5 negatives are mostly false negatives.

Query	Passage
effects of yeast on body	The Side Effects of Chemotherapy on the Body. Chemotherapy drugs are powerful enough to kill rapidly growing cancer cells, but they also can harm perfectly healthy cells, causing side effects throughout the body.
age not to take off shoes at airline security	Airline Identification Requirements. Airlines do not typically require identification from passengers under the age of 18, but there are exceptions. Children under the age of 2 may ride on a parent's lap without purchasing a ticket, but the airline will require identification, such as a birth certificate, to prove the child's date of birth . . .
actress who plays alice on the magicians	Then portrayed as the animated Alice's real life counterpart by actress Mia Wasikowska as a more mature, grown up Alice in Disney's 2010 semi-sequel, live action/CGI film Alice and Wonderland Directed by Tim Burton. In the Broadway musical version, she will be played by Taylor Louderman . . .
can you absorb metals from plants	Answer 1: Photosynthesis is the ability of plants to absorb the energy of light, and convert it into energy for the plant. To do this, plants have pigment molecules which absorb the energy of light very well. The pigment responsible for most light-harvesting by plants is chlorophyll, a green pigment.The green color indicates that it is absorbing all the non-green light– the blues (425-450 nm), the reds and yellows (600-700 nm).he pigment responsible for most light-harvesting by plants is chlor . . .

Table 6. Sample of (query, passage) pairs from MSMARCO-Passage dev set with largest discrepancy between the CA and DE model scores. In most of these cases, the passage is not relevant to the query; however, there is a high degree of token overlap between the two, indicating superficial similarity.

C.6. Results with ColBERT scorer

Table 7 presents results using the ColBERT scorer. We find that with distillation using the softmax CE loss ((1)), we can *exceed* the performance of the cross-attention teacher model for the re-ranking task. This further indicates the viability of dual-encoder models for neural ranking.

	MSMARCO rerank		TREC DL19	
	MRR@10	**nDCG@10**	**MRR@10**	**nDCG@10**
Baselines: one-hot				
Cross-attention BERT	0.370	0.430	0.829	0.749
Dual-encoder ColBERT	0.356	0.416	0.839	0.703
Distilled ColBERT: prior work				
MSE (Hofstätter et al., 2020a)	$0.365^{\dagger}$	$0.428^{\dagger}$	—	—
Margin-MSE (Hofstätter et al., 2020a)	$0.370^{\dagger}$	$0.431^{\dagger}$	$0.862^{\dagger}$	$0.738^{\dagger}$
Distilled ColBERT: this work				
$\mathrm{M}^3\mathrm{SE}$	0.371	0.430	0.875	0.726
Softmax CE	0.376	0.437	0.835	0.737

Table 7. Results on MSMARCO Passage dev set with ColBERT model.

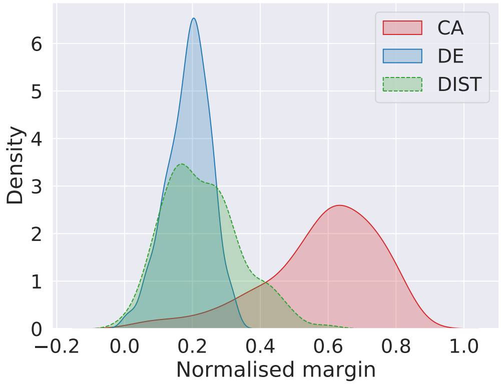

Figure 13. M^{3}SE distillation model margins on the MSMARCO-Passage test set. Compared to the DE model trained with one-hot labels, the distilled model is seen to more confidently distinguish positives from negatives, evidenced by the margin distribution having more mass on larger values.

Model	Supervision	Loss	MRR@10	nDCG@10
Cross-attention	One-hot	Softmax CE	0.370	0.430
Dual-encoder	One-hot	Softmax CE	0.249	0.299
Dual-encoder	Teacher	MSE	0.180	0.232
	Teacher	Margin MSE	0.251	0.299
	Teacher	KL	0.273	0.321
	Teacher	M^{3}SE	0.261	0.308

Table 8. Re-ranking results on MSMARCO-Passage dev set, using a small-bert-6-128 model.

D. Additional experiments: impact of distillation

D.1. Impact of distillation: effect on score distributions

We have previously seen in Figure 8 that our proposed M^{3}SE distillation strategy can improve the distribution of scores for positives and negatives. Figure 13 shows the impact that such distillation has on the margins between positive and negatives, compared to standard training on one-hot labels. The distilled model is seen to more confidently distinguish positives from negatives, as evidenced by the margin distribution having more mass on larger values.

D.2. Impact of distillation: effect of student model architecture

The preceding distillation results have all employed a small-bert-6-768 model. It is of interest how sensitive the results are to this choice. To study this, we report in Table 8 the re-ranking results on MSMARCO-Passage when using a small-bert-6-128 model. We observe largely consistent trends as before: distillation using our proposed losses can close the gap between CA and DE model performance. Table 9 repeats this for a small-bert-2-768 model.

D.3. Impact of distillation: effect of number of documents in supervision

The results reported in the body employed a total of 20 documents per query for the softmax CE and M^{3}SE methods. As this involves significantly more information than the triplet data used, e.g., in training the margin MSE loss, it is of interest to tease apart how much of the gains from the method come from this increased supervision, versus the loss itself. Table 10 studies the effect of varying the number of documents. We see that there are diminishing returns beyond a certain point: using 100 documents yields qualitatively similar performance to using 20 documents. At the same time, there is a marked difference in performance when using 2 documents versus 10.

Model	Supervision	Loss	MRR@10	nDCG@10
Cross-attention	One-hot	Softmax CE	0.370	0.430
Dual-encoder	One-hot	Softmax CE	0.281	0.331
Dual-encoder	Teacher	MSE	0.259	0.313
	Teacher	Margin MSE	0.313	0.368
	Teacher	KL	0.311	0.370
	Teacher	M^{3}SE	0.286	0.337

Table 9. Re-ranking results on MSMARCO-Passage dev set, using a small-bert-2-768 model.

# of documents	Loss	MRR@10	nDCG@10
2	Softmax CE	0.327	0.386
	M^{3}SE	0.323	0.379
10	Softmax CE	0.333	0.391
	M^{3}SE	0.326	0.382
20	Softmax CE	0.338	0.396
	M^{3}SE	0.349	0.406
100	Softmax CE	0.334	0.393
	M^{3}SE	0.325	0.381

Table 10. Sensitivity analysis of distillation losses to number of documents per sample.

E. Additional experiments: assorted

E.1. Impact of negative mining on re-ranking performance

We consider the value of adding additional negatives during DE model training. Following Karpukhin et al. (2020); Qu et al. (2021), we consider the use of within-batch (also referred to as in-batch) and uniform negatives. For the latter, in each training minibatch, we draw a sample of B_{uni} documents drawn uniformly at random from the entire set of documents; each such document is treated as a negative document in the loss. For the former, from each training minibatch $\{(q_i, \mathbf{d}_i, \mathbf{y}_i)\}_{i=1}^{B}$, we compute the set of all observed documents $\mathcal{D}_{\mathrm{obs}} = \cup_{i=1}^{B} \cup_{k=1}^{K} \{d_{ik}\}$. For each sample $(q_i, \mathbf{d}_i, \mathbf{y}_i)$, we then use each element in $\mathcal{D}_{\mathrm{obs}} - \cup_{k=1}^{K}\{d_{ik}\}$ as a negative document in the loss.

Table 11 summarises the results for the MSMARCO re-ranking task. We find that adding these negatives has a small gain for the one-hot model performance. However, for the distilled objectives, there is limited gain (and sometimes even a degradation) in performance. More study of this issue is warranted, but one hypothesis is that the re-ranking task is inherently concerned with ranking the outputs of a BM25 model. These outputs are precisely used to construct the training data used for distillation. It is thus possible that adding additional negatives — which are unlikely to appear as candidates for re-ranking — do not bring significantly useful information.

We emphasise here also that the results reported are for the *re-ranking*, as opposed to *retrieval* task. For the latter, adding within-batch and uniform negatives is intuitively and empirically valuable, as noted in Karpukhin et al. (2020); Qu et al. (2021). For example, in training the one-hot DE model on the triplet data, the retrieval MRR@10 is 0.281. When adding within-batch and uniform negatives, this increases to 0.314. Given that the re-ranking MRR@10 remains relatively unchanged (0.312 versus 0.310), this is further indication of the re-ranking and retrieval objectives not being perfectly aligned.

Supervision	Loss	MRR@10
One-hot	Softmax CE	0.310
	+ negative mining	0.312
Teacher	Margin MSE	0.334
	+ negative mining	0.335
Teacher	M^3SE	0.349
	+ negative mining	0.324
Teacher	KL	0.346
	+ negative mining	0.342

Table 11. Results of negative mining for DE models, on MSMARCO passage re-ranking task.

E.2. Full retrieval performance

Thus far, we have focussed on the re-ranking performance of all models, wherein the predictions from a BM25 retriever are provided as input. This allows for the comparison of DE and CA models on an equal footing. One may however naturally wonder how the proposed methods fare in the retrieval phase, wherein the models score *all* possible passages. Table 12 summarises the results in this setting on MSMARCO. We again see that the proposed distillation techniques offer strong gains over standard one-hot training of the DE model, as well as existing distillation techniques. Note that the poor retrieval performance of the MSE loss is a result of the model strongly overfitting to the teacher scores on the provided documents.

It is also of interest to analyse the generalization gap issue discussed in Section 3.2 in relation to dual encoder retrievers. Note that it is not possible to compare the retrieval performance of CA and DE models directly, since CA models do not support efficient nearest neighbor search. We therefore analyze the train-test gap for DE models during retrieval. We used a DE model trained on MSMARCO, employing uniformly sampled negatives, akin to the ANCE method (Xiong et al., 2021). Here, the test MRR@10 when performing full retrieval is 0.314, while the train MRR@10 is 0.364. This illustrates that the DE retrievers too can have a significant generalization gap.

Recent works such as Qu et al. (2021), Ruiyang Ren & Wen (2021), Keshav Santhanam (2022) propose elegant hard-negative mining schemes to significantly improve the performance of DE models for retrieval. Furthermore, Ni et al. (2021) shows that using larger models can also boost the generalization of dual-encoder retrievers. Exploring these is of interest, though we re-iterate that our primary goal is not on advancing the state-of-the-art for the MSMARCO dataset.

E.3. Analysis of TREC DL19 predictions

Table 1 suggests a notable gap in performance for our proposed RankDistil variant and the margin MSE loss (Hofstätter et al., 2020a). Here, we take a closer look at the loss cases for our method. The TREC DL19 data comprises a total of 43 queries, each with between ~ 100 to 500 rated passages. Of these queries, the MRR@10 of the DE models trained with margin MSE and softmax CE agree on 10. On the queries with disagreement, 6 cases favour margin MSE, and 4 cases favour M^3SE.

For the 6 queries where M^3SE ostensibly underperforms, we inspect the top-5 scoring passages for both margin MSE and softmax CE in Table 13 and 14. The cells shaded blue correspond to passages rated positive. Several other passages are however seen to be equally valid answers to the source query. Indeed, we submit that in *all* cases, the predictions from RankDistil are of at least the same quality as the margin MSE.

Model	**MRR**	**nDCG**
Baselines: one-hot		
BM25 (Robertson & Zaragoza, 2009)	0.194†	0.241†
ANCE (Xiong et al., 2021)	0.330†	—
Cross-attention BERT (12-layer)	N/A	N/A
Dual-encoder BERT (6-layer)	0.281	0.331
Distilled dual-encoder: prior work		
MSE (Hofstätter et al., 2020a)	0.000*	0.000*
Margin MSE (Hofstätter et al., 2020a)	0.319	0.375
RankDistil-B (Reddi et al., 2021)	0.000*	0.000*
Distilled dual-encoder: this work		
M^{3}SE (4)	0.337	0.394
Softmax CE (1)	0.334	0.392

Table 12. Summary of full retrieval MRR@10 and nDCG@10 for all methods on MSMARCO Passage. We compare cross-attention, dual-encoder, and distilled dual-encoder BERT models. We highlight the best performing DE based model. Distilling the dual-encoder with our proposed techniques significantly improves performance over one-hot training and existing distillation techniques. Results marked † are quoted from the corresponding reference, "N/A" are not applicable (e.g., the cross-attention model is not feasible to apply for retrieval), and "—" are not available from the reference. * See text for discussion.

Query	**Top scoring passages**	
	Margin MSE	**M^{3}SE**
who is robert gray	Who is Henry Gray? Henry Gray is an African-American blues piano player and singer. He has been play. . .	Robert Grey (born 21 April 1951 in Marefield, Leicestershire) is an English musician best known as t. . .
	Robert Gray was the Democratic candidate for governor of Mississippi in the 2015 elections. Gray won. . .	Who is Henry Gray? Henry Gray is an African-American blues piano player and singer. He has been play. . .
	Kenneth Gray (I) Kenneth Gray is an actor, known for Love, Lies and Murder (1991), Retribution (1987. . .	Robert Gray was the Democratic candidate for governor of Mississippi in the 2015 elections. Gray won. . .
	Kenneth Gray (I) Actor. Kenneth Gray is an actor, known for Love, Lies and Murder (1991), Retributio. . .	Robert Gray. A surprise came on the Democratic side in the race for Mississippi Governor. Robert Gra. . .
	Robert Grey (born 21 April 1951 in Marefield, Leicestershire) is an English musician best known as t. . .	William Thomas Gray. Billy Gray was born on January 13, 1938 in Los Angeles, California, USA as Will. . .
define visceral?	Definition of visceral. 1 : felt in or as if in the internal organs of the body : deep a visceral co. . .	Definition of Visceral. Visceral: Referring to the viscera, the internal organs of the body, specifi. . .
	Definition of Visceral. Visceral: Referring to the viscera, the internal organs of the body, specifi. . .	Definition of Visceral. Visceral: Referring to the viscera, the internal organs of the body, specifi. . .
	Definition of visceral. 1 1 : felt in or as if in the internal organs of the body : deep a visceral. . .	Medical Definition of Visceral. Visceral: Referring to the viscera, the internal organs of the body,. . .
	Definition of Visceral. Visceral: Referring to the viscera, the internal organs of the body, specifi. . .	Definition of visceral. 1 : felt in or as if in the internal organs of the body : deep a visceral co. . .
	Definition of visceral. 1 1 : felt in or as if in the internal organs of the body : deep a visceral. . .	Define visceral: felt in or as if in the internal organs of the body : deep; not intellectual : inst. . .
what is the daily life of thai people	T he following concepts are part of Thai everyday life: or JAI YEN is more a way . . .	The Daily Life of a Thai Monk The Sangha World in Thailand consists of about 200,000 monks and 85,00. . .
	The following concepts are part of Thai everyday life: or JAI YEN is more a way o. . .	An important thing in everyday life is SANUK. Thai people love to have fun together. SANUK can repre. . .
	The population of Thailand is approximately 67.5 million people, with an annual growth rate of about. . .	T he following concepts are part of Thai everyday life: or JAI YEN is more a way . . .
	For the rapcore band, see Every Day Life. Everyday life or Daily life or Routine life is a phrase us. . .	The following concepts are part of Thai everyday life: or JAI YEN is more a way o. . .
	Everyday life. Everyday life, daily life or routine life comprises the ways in which people typicall. . .	The population of Thailand is approximately 67.5 million people, with an annual growth rate of about. . .

Table 13. Comparison of top-5 scoring passages for margin MSE and softmax CE on TREC DL19 test set. The cells shaded blue correspond to passages rated positive. Several other passages are however seen to be equally valid answers to the source query.

Query	**Top scoring passages**	
	Margin MSE	**M^{3}SE**
what are the three percenters?	The Three Percenters, formed in late 2008, are a loosely organized movement centered around an obscu. . .	Try to understand that being a Three Percenter (Threeper, 3%, 3 Percenter, etc) is more of an idea t. . .
	III% Club Merchandise & III% Logo Gear. The Three Percenters Club official merchandise is designed a. . .	But this still doesnt answer the question. So how do you know if you are a Three Percenter? Try t. . .
	Rhodes has written supportively of the Three Percenters, while at least two participants carried the. . .	So let me see if I can help. In fact, let me provide you with well over 50 ways to tell whether or n. . .
	Try to understand that being a Three Percenter (Threeper, 3%, 3 Percenter, etc) is more of an idea t. . .	The Three Percenters, formed in late 2008, are a loosely organized movement centered around an obscu. . .
	But this still doesnt answer the question. So how do you know if you are a Three Percenter? Try t. . .	III% Club Merchandise & III% Logo Gear. The Three Percenters Club official merchandise is designed a. . .
how are some sharks warm blooded	Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblood. . .	Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblood. . .
	Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblood. . .	Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblood. . .
	Are White Sharks warm-blooded or cold-blooded? White sharks are part of the fish family, so they mus. . .	Are White Sharks warm-blooded or cold-blooded? White sharks are part of the fish family, so they mus. . .
	Great white sharks are some of the only warm blooded sharks. This allows them to swim in colder wate. . .	Great white sharks are some of the only warm blooded sharks. This allows them to swim in colder wate. . .
	These sharks can raise their temperature about the temperature of the water; they need to have oc. . .	The Salmon Shark is one of the warmest of the warm-bodied sharks. Measurements of its epaxial (upper. . .
what are the social determinants of health	The social determinants of health are the circumstances in which people are born, grow up, live, wor. . .	Social determinants of health reflect the social factors and physical conditions of the environment . . .
	Social determinants of health reflect the social factors and physical conditions of the environment . . .	Social determinants of health are economic and social conditions that influence the health of people. . .
	The social determinants of health are linked to the economic and social conditions and their distrib. . .	Social determinants of health are conditions in the environments in which people are born, live, lea. . .
	Social determinants of health are conditions in the environments in which people are born, live, lea. . .	Back to Top. Social determinants of health are conditions in the environments in which people are bo. . .
	Determinants of health are factors that contribute to a person's current state of health. These fact. . .	The social determinants of health are the circumstances in which people are born, grow up, live, wor. . .

Table 14. Comparison of top-5 scoring passages for margin MSE and softmax CE on TREC DL19 test set. The cells shaded blue correspond to passages rated positive. Several other passages are however seen to be equally valid answers to the source query.