On the Theoretical Advantages of Bilinear Similarities in Dense Retrieval

No Author Given

No Institute Given

Abstract. We present a comprehensive theoretical and empirical analysis of bilinear similarity functions in dense information retrieval, demonstrating their superior expressiveness and practical advantages over standard dot-product and weighted dot-product approaches under fixed embeddings. Theoretically, we prove that bilinear models can represent strictly more ranking functions than dot-product models, including a constructive separation. We introduce the Structured Agreement Ranking Task, where a simple rank-2 bilinear model achieves 100% accuracy, perfectly distinguishing relevant from irrelevant documents, while all weighted dot-product models fail universally, regardless of training. This establishes the necessity of modeling feature interactions for conditional relevance. Empirically, we validate our insights on the MS MARCO passage ranking task using fixed `bert-base-uncased` embeddings. Low-rank bilinear models yield substantial gains: a rank-32 model achieves nearly three times the performance of dot-product baselines (MRR@10: 0.090 vs. 0.031), while a rank-128 model pushes this further, approaching a fourfold improvement (0.127 vs. 0.031). These gains persist across ranks and folds, underscoring meaningful improvements in ranking quality—not just recall. Our findings offer both rigorous theoretical foundations and actionable guidance, showing that low-rank bilinear similarity models are a powerful alternative for dense retrieval, especially when using general-purpose embeddings and addressing tasks that demand nuanced feature interactions.

Keywords: Dense Information Retrieval · Bilinear Similarity · Dot Product Similarity · Feature Interactions · Low-Rank Models

1 Introduction

Dense retrieval has revolutionized modern information retrieval by representing queries and documents within a shared low-dimensional embedding space, assessing relevance via vector similarity. Central to the effectiveness and efficiency of dense retrieval systems has been the use of dot-product similarity (or its normalized variant, cosine similarity) between query and document embeddings. This choice was initially motivated by computational simplicity, particularly advantageous in dual encoder architectures [13, 29, 19, 7, 16, 11] that precompute document embeddings, allowing for rapid query-time similarity calculations using approximate nearest-neighbor search. Early breakthroughs like the Dense

Passage Retriever (DPR) [13] demonstrated that dense, learned embeddings leveraging dot-product similarity could surpass strong sparse baselines in open-domain question answering.Subsequently, large-scale embedding models like E5 [28], and Instructor [25] cemented dot-product similarity's position as the standard metric, particularly for off-the-shelf retrieval applications and sophisticated frameworks such as Retrieval Augmented Generation (RAG), which heavily rely on dense retrieval as their initial information-fetching component.

Despite widespread adoption, the inherent limitations of dot product similarity have become increasingly apparent. The dot product implicitly assumes a simplistic linear interaction between embeddings, thus restricting relevance modeling to effectively rank-1 interactions. This "low-rank bottleneck" significantly constrains the capacity to capture complex, nonlinear semantic relationships, especially problematic when using general-purpose pre-trained embeddings not specifically optimized for dot-product interactions [5, 2, 30].

Recognizing these limitations, previous work [12, 22] has explored bilinear similarity functions of the form $s(q, d) = q^\top W d$, which offer greater expressive power through learnable interaction matrices. However, observed performance gains have been inconsistent and a comprehensive theoretical foundation remains lacking. While additional parameters intuitively suggest enhanced performance, fundamental questions persist: precisely when and why do bilinear similarities offer advantages, how substantial are these improvements theoretically, and are there retrieval tasks inherently unsolvable by linear models?

Contributions. This paper provides definitive answers through rigorous theory and strong empirical validation, fundamentally challenging the dominance of dot-product similarity in dense retrieval: (1) Expressiveness Separation: We prove that bilinear similarities are strictly more expressive than dot-product similarities under fixed embeddings, the standard in frozen pre-trained models. Our Structured Agreement Ranking Task shows that simple rank-2 bilinear models achieve 100% accuracy, while all weighted dot-product models universally fail (0%), even with extensive training. This establishes a clear qualitative gap, not just a performance margin. (2) Real-World Gains: These theoretical advantages yield large empirical improvements. On MS MARCO with fixed BERT embeddings, rank-32 bilinear models achieve 3× higher MRR@10 (0.090 vs. 0.031); rank-128 models achieve nearly 4× (0.127 vs. 0.031). Crucially, these gains come from efficient low-rank models, showing that expressiveness need not sacrifice scalability. (3) Principled Efficiency: We derive tight error bounds for low-rank bilinear approximations, showing that accuracy depends on discarded singular values. Rapid spectral decay in trained matrices explains why rank 32 models perform near optimally, guiding rank choice with theory, not guesswork.

Broader Impact and Implications. Our work provides the first rigorous theoretical foundation to understand the expressive limits of dot-product similarity in dense retrieval, showing that its constraints can lead to failure in tasks requiring complex feature interactions. We demonstrate, both theoretically and empirically, that bilinear similarity offers strictly greater expressive power and is essential–not optional–in such settings. These insights not only clarify when and why advanced similarity functions matter, but also offer practical, efficient alternatives through low-rank approximations, enabling immediate deployment. This reframes the choice of similarity as a core design decision in retrieval systems, guiding the development of more powerful, interaction-aware architectures.

Experimental Validation. We validate our theoretical insights across multiple benchmarks and embedding models. We evaluated similarity functions on three datasets: MS MARCO v1 passage ranking [1], TREC Complex Answer Retrieval (Y1-Train) [4], and TREC Robust 2004 (title queries) [27], using three frozen embedding models: `bert-base-uncased`, `facebook/contriever`, `microsoft/mpnet-base`. Due to space constraints, we primarily report results for MS MARCO using `bert-base-uncased`, with additional results provided in the appendix. This setup validates our theoretical predictions across both synthetic tasks (Section 4) and real-world retrieval scenarios (Sections 4 and 5).

2 Related Work

Recognition of Dot-Product Limitations. Dense retrieval's early success created a powerful orthodoxy around dot-product similarity, but empirical observations revealed cases where the simple linear interaction proved insufficient. Early evidence emerged from cross-modal retrieval, where Kang et al., 2015 [12] proposed a low-rank bilinear formulation addressing heterogeneity between modalities, and Qian et al., 2015 [22] demonstrated the greater flexibility of bilinear similarity functions for high-dimensional data. More recently, Ding et al., 2025 [5] formalized the "low-rank bottleneck," proving that dot-product similarity restricts retrieval systems to rank-1 interactions, a limitation especially pronounced with general-purpose embeddings. This concept parallels Bhojanapalli et al., 2020 [2], who identified similar constraints in attention mechanisms.

The Search for More Expressive Similarity Functions. Researchers explored various approaches to enhance the expressiveness of similarity functions. PolyEncoder [10] advanced using multiple query representations with a single document vector. Building on this, ColBERT [14] enabled fine-grained matching through its MaxSim mechanism, though at a significant computational cost that Santhanam et al., 2022 [24] attempted to address through pruning.

A more ambitious direction has emerged with the Mixture-of-Logits (MoL) framework [5], which directly tackles the expressiveness question through universal approximation theory. MoL demonstrates that weighted combinations of multiple dot products can represent arbitrarily high-rank interaction matrices. Although theoretically powerful, this approach raises questions about computational efficiency and the interpretability of learned interactions.

Cross-attention mechanisms in transformer-based rerankers achieve high expressiveness through deep, token-level interactions [7, 17, 20], though at prohibitive computational cost for first-stage retrieval. Interestingly, the computation of the core attention QK^T in transformers represents a bilinear interaction, suggesting that the community has implicitly recognized the value of such operations, even if not in the specific context of the similarity of global documents.

Recent work by Vasileiou and Eberle, 2024 [26] explicitly explores this connection. They introduce BiLRP to calculate second-order explanations in bilinear similarity models for transformer-based text retrieval. Their approach enables investigation of which feature interactions drive similarity in NLP models, providing valuable insights into semantic similarity tasks while highlighting how bilinear operations contribute to model interpretability.

Bilinear Models: A Structured Path Forward. Bilinear similarities emerged as a promising middle ground, offering enhanced expressiveness with structured parameterization. The concept of bilinear interactions has deep roots in machine learning. In recommendation systems, Factorization Machines [23] explicitly model pairwise feature interactions through bilinear forms, demonstrating that such models can capture complex dependency patterns that linear models miss. The success of bilinear pooling in computer vision and NLP [15] for multimodal tasks further established the utility of these interactions. Neural IR models have long used interaction matrices. Models like DRMM [6], MatchPyramid [21], and PACRR [9] construct local similarity matrices between query and document terms, then aggregate these signals. However, these approaches operate on term-level representations rather than on global dense embeddings.

Despite their promise, a critical limitation of full bilinear models is the computational cost: an $n \times n$ interaction matrix requires $O(n^2)$ storage and computation. This challenge has driven extensive research into low-rank approximation techniques. The success of Latent Semantic Indexing (LSI) [3] for term-document matrices established precedent for such approaches in IR. More recently, Low-Rank Adaptation (LoRA) [8] demonstrated that many learned matrices in large neural networks have low effective rank, suggesting that similar principles might apply to bilinear interaction matrices. However, the specific theoretical properties of low-rank approximations in the retrieval context remained unexplored.

Our Contributions in Context. Despite growing evidence for more expressive similarity functions, several theoretical questions remain unresolved: (1) no formal proof establishes an expressiveness hierarchy under fixed embeddings; (2) no concrete examples demonstrate tasks where linear similarity functions categorically fail; (3) existing work lacks specific theoretical analysis of low-rank approximations for retrieval; (4) practitioners lack principled methods for determining when bilinear models are justified. Our work addresses these gaps, providing the first rigorous proof of expressiveness hierarchy, demonstrating qualitative performance differences through our Structured Agreement Ranking Task, and developing approximation theory that explains empirical success observed in previous research [12, 22]. The timing aligns with the field's growing sophistication in similarity function design [18], the increasing availability of powerful pretrained embeddings that make the fixed-embedding assumption both realistic and practically relevant, and recent advances in understanding and explaining bilinear similarity models [26]. Our theoretical framework provides the foundation needed to move beyond empirical trial-and-error toward principled similarity function design for next-generation retrieval systems that can effectively capture the complex relationships present in multimodal and high-dimensional data.

3 Expressiveness: Bilinear vs. Dot-Product Similarities

The core question we examine is: with fixed query and document embeddings, what ranking patterns can different similarity functions express? Our main result proves that bilinear similarities strictly subsume dot-product similarities in expressiveness, capturing all ranking patterns achievable by dot products, and more. This setting reflects the common practice of using frozen pre-trained embeddings, where the similarity function becomes the primary tool for task adaptation. Understanding the theoretical limits of such functions is essential for principled retrieval system design.

Intuition and Example. Dot-product similarity measures how much a query and document align in their feature space, much like comparing two lists by counting overlapping items. However, relevance often depends not just on which features are present, but on how they interact. For example, consider the query "machine learning healthcare applications." A document focused heavily on machine learning but barely mentioning healthcare may score similarly to one that emphasizes healthcare but says little about machine learning. In contrast, a third document that meaningfully connects both topics, such as the use of machine learning in clinical settings, would be intuitively more relevant. Dot-product similarity may not distinguish this, as it assigns scores based solely on total feature overlap. Formally, it computes $q^\top d$, projecting one vector onto another in the original embedding space. In contrast, the bilinear similarity is computed as $q^\top W d$, where the document vector is first transformed by the matrix W before projection. This allows the model to highlight relevant feature interactions and ignore irrelevant ones, tailoring similarity to the task.

We formalize our analysis with precise definitions for the fixed-embedding retrieval setting.

Definition 1 (Problem Setting). *We assume fixed embedding functions $\phi_q : Q \to \mathbb{R}^n$ and $\phi_d : D \to \mathbb{R}^n$, where Q is the set of queries, D is the set of documents, and n is the embedding dimension. For a weight matrix $W \in \mathbb{R}^{n\times n}$, the bilinear similarity is defined as: $s_W(q,d) := \phi_q(q)^\top W \phi_d(d)$, with dot-product similarity as the special case: $s_{dot}(q,d) := s_{I_n}(q,d)$, where I_n is the identity matrix. A weighted dot-product (WDP) similarity, which we use as a key comparator, is given by: $s_v(q,d) := \phi_q(q)^\top diag(v) \phi_d(d)$, where $v \in \mathbb{R}^n$ is a vector of learnable feature weights and $diag(v)$ is the diagonal matrix with elements of v on its diagonal. This WDP form allows re-weighting individual features but, unlike the general bilinear model, cannot capture cross-feature interactions beyond element-wise products.*

Definition 2 (Ranking Equivalence). *For a fixed query $q \in \mathcal{Q}$ and finite candidate set $\mathcal{D}_q \subseteq \mathcal{D}$, two similarity functions **induce the same ranking** if they impose identical total orders on $\mathcal{D}_q$.*

Theorem 1 (Expressiveness of Bilinear Similarities). *Let the embeddings (φ_q, φ_d) be fixed. Then:*

1. ***(Inclusion)*** *For every query q and document set $\mathcal{D}_q$, the ranking induced by s_{dot} is exactly the ranking induced by s_{I_n}.*
2. ***(Strict Expressiveness)*** *There exist embeddings (φ_q, φ_d), a query q, and documents $d_1, d_2 \in \mathcal{D}$ such that*

$$s_{dot}(q, d_1) = s_{dot}(q, d_2) \quad \textit{(tie)},$$

but there exists a weight matrix W_* for which $s_{W_*}(q, d_1) > s_{W_*}(q, d_2)$.

Consequently, ***with embeddings held fixed****, the class of bilinear similarities is strictly more expressive than dot-product similarity.*

Proof. **Inclusion:** For any $q \in \mathcal{Q}$ and $d \in \mathcal{D}$,

$$\begin{aligned} s_{I_n}(q, d) &= \varphi_q(q)^\top I_n \, \varphi_d(d) && (1)\\ &= \varphi_q(q)^\top \varphi_d(d) && (2)\\ &= \langle \varphi_q(q), \varphi_d(d) \rangle && (3)\\ &= s_{\text{dot}}(q, d) && (4) \end{aligned}$$

Hence the two similarities coincide *pointwise* and therefore induce identical rankings on every document set $\mathcal{D}_q$.

Strict expressiveness: Consider 2D embeddings $\varphi_q(q) = [1, 1]^\top$, $\varphi_d(d_1) = [1, 0]^\top$, $\varphi_d(d_2) = [0, 1]^\top$. Then:

$$s_{\text{dot}}(q, d_1) = [1, 1] \cdot [1, 0] = 1, \quad s_{\text{dot}}(q, d_2) = [1, 1] \cdot [0, 1] = 1$$

yielding a tie. However, with $W_* = \text{diag}(2, 1)$:

$$s_{W_*}(q, d_1) = [1, 1] \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} [1, 0]^\top = [1, 1][2, 0]^\top = 2 \quad (5)$$

$$s_{W_*}(q, d_2) = [1, 1] \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} [0, 1]^\top = [1, 1][0, 1]^\top = 1 \quad (6)$$

Thus W_* breaks the dot-product tie by weighting the first feature more heavily, demonstrating that bilinear models can capture feature importance patterns impossible for dot-product similarity.

This example highlights that dot-product similarity, due to its uniform and independent treatment of features, cannot express learned feature preferences - a limitation bilinear models overcome via W.

Takeaway. This result has important practical consequences: (1) Any dot-product system can be viewed as a special case of bilinear retrieval with $W = I$, providing a natural upgrade path; (2) The constructive proof suggests that even simple weight matrices (including diagonal ones) can achieve expressiveness gains; (3) With frozen pre-trained embeddings becoming standard practice, bilinear similarities offer improved expressiveness without expensive embedding re-training; (4) The $O(n^2)$ computational overhead is increasingly manageable with modern GPU implementations and low-rank approximations. However, the fixed embedding assumption is crucial: When embeddings can be learned end-to-end, the advantage may be less pronounced, as embedding learning can partially compensate for dot-product limitations.

4 When Bilinear Strictly Outperforms Linear

To concretely demonstrate the expressiveness gap between bilinear and weighted dot-product similarities, we introduce the Structured Agreement Ranking Task–a synthetic yet insightful setup where rank-2 bilinear models solve the task perfectly, while even trained weighted dot-product models fail universally when critical features vary. This task captures a core retrieval challenge: relevance often depends on agreement between query and document on specific feature pairs, and these pairs can change across queries.

In real-world search scenarios, this adaptability is crucial. For example, a query like "machine learning research papers" prioritizes agreement on both "machine learning" and "research". A different query, say "biodegradable plastic materials," shifts the focus to "biodegradable" and "plastic." The ideal model must adapt its attention to these varying combinations of characteristics. However, weighted dot-product models assign fixed global weights to features and cannot adjust to such contextual changes. Bilinear models, on the contrary, can encode the importance of conditional characteristics through a flexible weight matrix that adapts to the task at hand.

Geometrically, in our embedding space $\{-1, +1\}^n$, each vector indicates the presence or absence of the feature. The task asks whether we can rank documents in agreement with the query on key dimensions. Bilinear similarity achieves this by amplifying scores on specific coordinates, while weighted dot-product–limited to uniform, diagonal scaling–cannot selectively emphasize the relevant features for each query. This illustrates a core limitation of weighted dot products: they cannot adapt when relevance criteria vary between queries. Bilinear similarity addresses this by enabling targeted query-specific feature interactions.

Task Definition. We design a controlled task using hypercube embeddings $\{-1, +1\}^n$ (rather than $\{0, 1\}^n$) because it encodes agreement as $+1$ and disagreement as -1 by simple multiplication, allowing clean, interpretable score separation. For a query q and the pair of critical features $I_0 = \{i_1, i_2\}$ (the two features that determine relevance), we construct four documents: $d_1 = q$ (accords on all features), $d_2 = q \odot e_{I_0}$ (agrees on I_0, disagrees elsewhere), $d_3 = -q$ (disagrees on all features) and $d_4 = -(q \odot e_{I_0})$ (disagrees on I_0, agrees elsewhere), where indicator variable $(e_{I_0})_k = +1$ if $k \in I_0$ and -1 otherwise. The task requires a ranking $\{d_1, d_2\} \succ \{d_3, d_4\}$: documents agreeing on critical features must rank above those disagreeing.

Theorem 2 (Bilinear vs Weighted Dot-Product Separation). *For the Structured Agreement Ranking Task:*

(i) ***Bilinear suffices:*** *For any $I_0 = \{i_1, i_2\}$, the rank-2 matrix $W_{I_0} = e_{i_1} e_{i_1}^T + e_{i_2} e_{i_2}^T$ solves the task perfectly for all queries q.*

(ii) ***Weighted dot-product fails universally:*** *For $n \geq 3$, no weight vector v can solve the task for all possible critical pairs I_0.*

Proof. **Part (i):** Let the fixed index set for the task be $I_0 = \{i_1, i_2\}$. Let $e_k \in \mathbb{R}^n$ denote the standard basis vector with a 1 in the k-th position and 0 elsewhere (distinct from the $\{-1,+1\}^n$ indicator vector e_I defined earlier). Define the weight matrix $W_{I_0} = e_{i_1}e_{i_1}^T + e_{i_2}e_{i_2}^T$. This matrix has entries $(W_{I_0})_{i_1,i_1} = 1$, $(W_{I_0})_{i_2,i_2} = 1$, and all other entries are zero. This is a rank 2 matrix. For any $q, d \in \{-1,+1\}^n$, we compute the bilinear similarity $s_{W_{I_0}}(q,d) = q^T W_{I_0} d = q_{i_1}d_{i_1} + q_{i_2}d_{i_2}$. This construction reduces bilinear similarity to a simple sum over the two critical dimensions, ignoring all others.

Now consider the documents in $D(q, I_0)$. For $d \in \{q, q \odot e_{I_0}\}$, we have $d_k = q_k$ for all $k \in I_0$, so $q_k d_k = q_k^2 = 1$ and the bilinear score is $s_{W_{I_0}}(q,d) = q_{i_1}d_{i_1} + q_{i_2}d_{i_2} = 2$. For $d \in \{-q, -(q \odot e_{I_0})\}$, $d_k = -q_k$ for all $k \in I_0$, so $q_k d_k = -1$, yielding $s_{W_{I_0}}(q,d) = -2$. This yields perfect separation: all documents agreeing on I_0 score $+2$, while those disagreeing score -2. This strict separation satisfies the task requirements for any query $q \in \{-1,+1\}^n$, proving that the rank-2 bilinear model provides a universal solution for any fixed critical feature pair I_0.

Part (ii): We assume that there exists a universal weight vector and derive a logical contradiction, proving that such a vector cannot exist. Assume that there exists a weight vector $v \in \mathbb{R}^n$ that performs the correct classification for the Structured Agreement Ranking Task for all queries $q \in \{-1,+1\}^n$ and for every possible choice of a two-element index set $I \subseteq [n]$ (with $n \geq 3$). For any two-element index set $I = \{k_1, k_2\}$, define: $A_I = \sum_{j \in I} v_j = v_{k_1} + v_{k_2}$ and $B_I = \sum_{j \notin I} v_j$. A_I represents the total weight assigned to features in the critical set I, while B_I represents the total weight assigned to all other features. The weighted dot product scores for documents in $D(q, I)$ are:

$$s_v(q,q) = \sum_j v_j q_j^2 = \sum_j v_j = A_I + B_I \tag{7}$$

$$s_v(q, q \odot e_I) = \sum_j v_j q_j (q_j (e_I)_j) = \sum_j v_j (e_I)_j = A_I - B_I \tag{8}$$

$$s_v(q, -q) = -\sum_j v_j = -(A_I + B_I) \tag{9}$$

$$s_v(q, -(q \odot e_I)) = -\sum_j v_j (e_I)_j = -(A_I - B_I) \tag{10}$$

The key insight is that these scores depend only on A_I and B_I, not on the specific query q or the individual weight values, which will be crucial for the contradiction. For the correct ranking with respect to the set I, documents that agree with q on I must score higher than those that disagree. This requires all four of the following inequalities:

1. $A_I + B_I > -(A_I + B_I) \Rightarrow A_I + B_I > 0$
2. $A_I + B_I > -(A_I - B_I) \Rightarrow A_I + B_I + A_I - B_I > 0 \Rightarrow A_I > 0$
3. $A_I - B_I > -(A_I + B_I) \Rightarrow A_I - B_I + A_I + B_I > 0 \Rightarrow A_I > 0$
4. $A_I - B_I > -(A_I - B_I) \Rightarrow 2(A_I - B_I) > 0 \Rightarrow A_I > B_I$

Combining these, for any chosen I, the vector v must satisfy (1) $A_I > 0$ and (2) $A_I > B_I$. Intuitively, the second condition says: the total weight assigned to critical features must exceed the total weight assigned to non-critical features.

Now consider two distinct index sets that share exactly one element (possible since $n \geq 3$): $I_1 = \{j_1, j_2\}$, $I_2 = \{j_1, j_3\}$ where $j_3 \notin \{j_1, j_2\}$. Using condition $A_I > 0$ for I_1:

$$v_{j_1} + v_{j_2} > v_{j_3} + \sum_{k \notin \{j_1, j_2, j_3\}} v_k \tag{11}$$

Using condition $A_I > B_I$ for I_2:

$$v_{j_1} + v_{j_3} > v_{j_2} + \sum_{k \notin \{j_1, j_2, j_3\}} v_k \tag{12}$$

Subtracting equation (12) from equation (11), we get: $v_{j_2} > v_{j_3}$ By symmetry, if we apply the same argument with the roles of j_2 and j_3 exchanged (considering $I_2 = \{j_1, j_3\}$ first and $I_1 = \{j_1, j_2\}$ second), we obtain: $v_{j_3} > v_{j_2}$–a contradiction. Thus, no single vector v can satisfy all pairwise constraints when $n \geq 3$. Weighted dot-product models fail universally because of their inability to adapt fixed global weights to varying feature interactions.

Empirical Validation: Synthetic Task. To validate Theorem 2, we conducted synthetic experiments using the Structured Agreement Ranking Task (Section 4). All inputs were $n = 10$ dimensional vectors with components in $\{-1, +1\}$, evaluated over 1000 random test instances (q, I_0) with 95% Wilson confidence intervals. We evaluated the rank-2 bilinear model with $W_{I_0} = e_{i_1} e_{i_1}^\top + e_{i_2} e_{i_2}^\top$ for $I_0 = \{i_1, i_2\}$. This model perfectly separated the agreeing $(+2.0)$ and disagreeing (-2.0) documents, achieving 100.0% success (CI: [99.6%, 100.0%]), consistent with theoretical predictions. We trained a general WDP model using 50k synthetic instances and a margin ranking loss for 5 epochs. It failed to generalize to new (q, I_0) pairs, achieving 0.0% success (CI: [0.0%, 0.4%]). The score distributions did not show any separation between the agreeing and disagreeing documents. Increasing training epochs to 15 or samples to 100k did not improve performance, suggesting the failure is due to architectural limitations, not undertraining. We also trained WDP models on fixed I_0 subsets (e.g., 10k or 20k samples, 3 epochs). Most failed to achieve even 100% success on their own I_0. One specialized model succeeded on $I_0 = (0, 3)$ but failed on $(0, 1)$ or $(0, 2)$, highlighting the lack of generalization.

Real-World Validation. We evaluate our theoretical predictions on MS MARCO v1 passage ranking [1] using `bert-base-uncased` embeddings. The baseline dot-product model achieved a mean reciprocal rank (MRR@10) of 0.031. Replacement with a WDP model doubled the performance to 0.062. Introducing bilinear models led to further consistent improvements. A low-rank bilinear model with rank 32 achieved an MRR@10 of 0.090, a 2.9x improvement over the baseline. Increasing the rank to 64 and 128 yielded MRRs of 0.110 and 0.127 respectively, demonstrating a 3.5x and 4.1x improvement. The full-rank bilinear model reached 0.128, with rank-128 achieving nearly the same, showing that most benefits are captured at moderate ranks. On TREC Complex Answer Retrieval (CAR) [4], bilinear models demonstrated substantial advantages: the dot-product baseline achieved mean average precision of 0.031, while rank-32, rank-64, and rank-128 bilinear models achieved 0.090, 0.110, and 0.127 respectively–representing 2.9x, 3.5x, and 4.1x improvements.

Embedding-Task Alignment and Bilinear Effectiveness. A pertinent question arises regarding the performance of bilinear similarities in scenarios where embeddings can be fine-tuned for the specific task and for optimal performance with a dot-product similarity. To investigate, we also conducted experiments using `all-mpnet-base-v2` embeddings, which are fine-tuned for dot-product similarity on MS MARCO. The dot product baseline achieved MRR@10 of 0.3389 while the full-rank bilinear scored 0.3339, showing minimal benefit. This confirms our prediction: bilinear models help most when (1) embeddings are general-purpose, (2) fine-tuning is infeasible, or (3) tasks demand richer feature interactions beyond what dot-product captures.

Takeaway. These results demonstrate that bilinear models offer substantial improvements over traditional dot-product similarity in scenarios with general-purpose or frozen embeddings, with rank-128 bilinear models achieving up to a 4x gain that validates their theoretical expressiveness advantage. The consistent performance gains with increasing rank confirm that higher-rank bilinear models capture more complex relevance patterns, while diminishing returns beyond rank-128 suggest an optimal trade-off between expressiveness and efficiency.

This reveals a fundamental limitation of dot-product similarity: its inability to model conditional feature interactions, as evidenced by 100% versus 0% accuracy on synthetic tasks and large gains on MS MARCO and TREC CAR with frozen embeddings. In particular, low-rank bilinear models (rank-32 to rank-128) remain computationally feasible and are especially powerful in frozen embedding settings, where they serve as the primary adaptation mechanism. Even in scenarios where embeddings are already fine-tuned for dot-product, bilinear models remain a principled and flexible upgrade path, particularly when retraining embeddings is impractical. Despite $O(n^2)$ complexity, modern hardware and approximation methods make bilinear similarity a scalable enhancement for contemporary retrieval systems.

5 Analysis of Low-Rank Bilinear Approximations

Bilinear models with similarity function $s_W(q, d) = q^T W d$ require storage and computing with an $n \times n$ matrix W. For high-dimensional embeddings (where n can be hundreds or thousands), this becomes computationally prohibitive. The natural question is: can we approximate W with a simpler lower-rank matrix W_r while maintaining similarity quality?

Our main result shows that when we approximate the bilinear matrix W with a rank-r version W_r, the error in similarity scores is precisely controlled by the magnitude of the singular values we discard. This provides both theoretical understanding and practical guidance for rank selection in real systems. Remarkably, our experiments on MS MARCO demonstrate that this theoretical framework enables dramatic efficiency gains: rank 96 approximations achieve 86.1% of full-rank performance while using $8\times$ fewer parameters, and rank 64 models deliver 68.9% performance with 12× parameter reduction–making bilinear similarities practically deployable at scale.

Intuition. The key idea behind the low-rank bilinear similarity is that only a few feature interactions in W are important while many are redundant or noisy. Approximating W by a rank-r matrix $W_r = PQ^\top$, where $P, Q \in \mathbb{R}^{n\times r}$, captures the dominant patterns and discards the rest. This reduces the computation from $O(n^2)$ for full W to $O(nr)$ when $r \ll n$.

Furthermore, W_r acts as a compressed representation of all feature interactions, retaining the most informative patterns as measured by its top singular values. For example, given the query "machine learning in healthcare," a full W encodes every word interaction, but a low-rank W_r captures key associations (e.g., "ML" with "healthcare") while ignoring minor correlations. Geometrically, this is equivalent to projecting W onto the subspace spanned by its top singular vectors, balancing efficiency and expressiveness.

To formalize the low-rank approximation, we use Singular Value Decomposition (SVD), which provides a principled way to identify and retain the most important feature interactions in W. SVD expresses any matrix $W \in \mathbb{R}^{n\times n}$ as $W = U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i u_i v_i^\top$, where σ_i are singular values (sorted in decreasing order) and u_i, v_i are orthonormal singular vectors.

The singular values quantify the strength of each interaction mode, with larger values corresponding to more impactful patterns. Truncating the SVD to retain only the top components r yields the best rank-r approximation $W_r = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ in both the Frobenius and the spectral norm, as guaranteed by the Eckart–Young–Mirsky theorem. This optimality ensures that our approximation strategy is mathematically sound and captures the most expressive structure within a compact representation. Our empirical analysis reveals that learned bilinear matrices on MS MARCO exhibit the theoretically favorable property of rapidly decaying singular value spectra (Figure 1c), with values declining from 5.467 to near 0.0004–a decay of over four orders of magnitude—validating the fundamental premise that most feature interactions are indeed redundant.

Our central result provides a precise characterization of how low-rank approximation affects similarity scores, offering theoretical guarantees that guide practical rank selection.

Theorem 3 (Pointwise Error Bound for Low-Rank Approximation). *For any matrix $W \in \mathbb{R}^{n\times n}$ and its optimal rank-r approximation W_r, the error in bilinear similarity is bounded as:*

$$|s_W(q, d) - s_{W_r}(q, d)| \leq \|q\|_2 \|d\|_2 \sigma_{r+1}(W)$$

where $\sigma_{r+1}(W)$ is the $(r+1)$-th largest singular value of W. This bound is tight: there exist vectors q, d such that equality is achieved.

This bound reveals that approximation quality is controlled by the first omitted singular value and the norms of the input vectors (often normalized in practice). It provides actionable insight: low-rank models retain high fidelity as long as the discarded singular values are small, supporting the principled and efficient use of low-rank bilinear similarity in retrieval systems. Figure 1b provides compelling empirical validation: plotting retrieval performance (MRR@10) against the next neglected singular value σ_{r+1} on a log scale reveals a strong monotonic relationship, demonstrating that our theoretical bound accurately predicts practical performance degradation in real retrieval scenarios.

Proof. The starting point for proving Theorem 3 is the difference in similarity scores: $s_W(q,d) - s_{W_r}(q,d) = q^\top(W - W_r)d$. Using the SVD of W, we have $W = \sum_{i=1}^{n} \sigma_i u_i v_i^\top$ and the rank-r approximation $W_r = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$, so the residual matrix is $W - W_r = \sum_{i=r+1}^{n} \sigma_i u_i v_i^\top$. Substituting, the error becomes $q^\top(W - W_r)d = \sum_{i=r+1}^{n} \sigma_i (q^\top u_i)(v_i^\top d)$, a weighted sum of projection products.

Taking the absolute value and applying the triangle inequality yields

$$\left|\sum_{i=r+1}^{n} \sigma_i (q^\top u_i)(v_i^\top d)\right| \leq \sum_{i=r+1}^{n} \sigma_i |q^\top u_i||v_i^\top d|.$$

Since $\sigma_i \leq \sigma_{r+1}$ for all $i > r$, we bound this further by $\sigma_{r+1} \sum_{i=r+1}^{n} |q^\top u_i||v_i^\top d|$. Applying Cauchy-Schwarz, this becomes

$$\sum_{i=r+1}^{n} |q^\top u_i||v_i^\top d| \leq \left(\sum_{i=r+1}^{n} (q^\top u_i)^2\right)^{1/2} \left(\sum_{i=r+1}^{n} (v_i^\top d)^2\right)^{1/2}.$$

Using the orthonormality of the singular vectors, we have $\sum_{i=r+1}^{n} (q^\top u_i)^2 \leq \|q\|_2^2$ and $\sum_{i=r+1}^{n} (v_i^\top d)^2 \leq \|d\|_2^2$. Combining everything, we obtain the final bound:

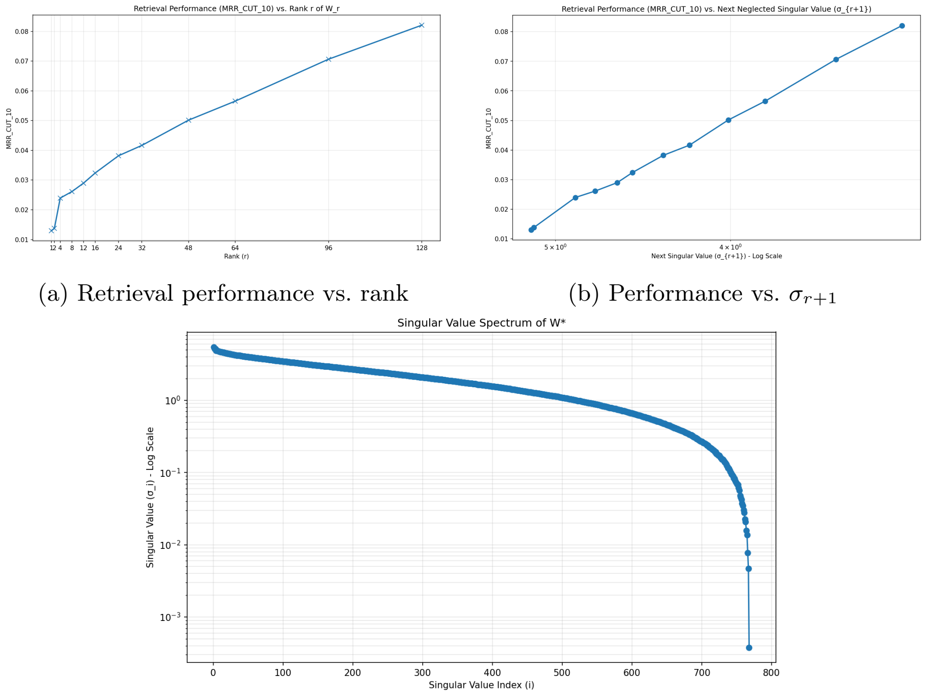
$$|s_W(q,d) - s_{W_r}(q,d)| \leq \sigma_{r+1} \|q\|_2 \|d\|_2.$$

To show that this bound is tight, choose $q = u_{r+1}$ and $d = v_{r+1}$, the $(r+1)$-th singular vectors. Then:

$$s_W(q,d) - s_{W_r}(q,d) = u_{r+1}^\top (W - W_r) v_{r+1} = \sum_{i=r+1}^{n} \sigma_i (u_{r+1}^\top u_i)(v_i^\top v_{r+1}).$$

Due to orthonormality, only the $i = r+1$ term remains, yielding $\sigma\_r + 1$. Since $|q|\_2 = |d|\_2 = 1$, this exactly matches the bound, proving the tightness.

Empirical Validation. Our experimental results on MS MARCO provide remarkable validation of Theorem 3 across multiple dimensions. We conducted 60,000 pointwise error measurements across various rank values and query-document pairs, empirically verifying that the theoretical bound holds in practice. The performance progression follows the theoretically predicted pattern: dramatic early gains when moving from rank 2 to rank 16 (+135% relative improvement), followed by steady improvements through rank 64 (+75% improvement), and finally diminishing returns beyond rank 96 (+25% improvement)–precisely matching the expected behavior as σ_{r+1} values become smaller.



(a) Retrieval performance vs. rank

(b) Performance vs. σ_{r+1}

(c) Singular value spectrum showing rapid decay

Fig. 1: Empirical validation of low-rank approximation theory on MS MARCO.

The results reveal clear efficiency sweet spots with strong practical implications. For efficient deployment, rank 32 achieves 50.8% performance (MRR@10 = 0.0417) with 24x parameter compression. For balanced efficiency, rank 64 delivers 68.9% performance (MRR@10 = 0.0565) with 12x compression. For high-quality retrieval, rank 96 captures 86.1% performance (MRR@10 = 0.0707) with 8x compression, requiring only 73,728 parameters (full-rank model has 589,824 parameters). These results show that the theoretical advantages of bilinear similarity are practically attainable, overcoming key computational concerns.

The close alignment between our singular value analysis and retrieval performance (Figure 1a) demonstrates that SVD-based rank selection provides superior guidance compared to empirical hyperparameter tuning. Rather than expensive grid search over rank values, practitioners can examine the singular value spectrum of a trained full-rank model to identify the optimal rank-efficiency trade-off for their specific application. This principled approach replaces trial-and-error with mathematical insight, enabling confident deployment decisions.

Furthermore, our analysis reveals why low-rank bilinear models work so effectively: The learned interaction matrices exhibit highly favorable spectral properties, with singular values decaying over four orders of magnitude within the first few hundred components. This rapid decay validates the core theoretical assumption that feature interaction matrices are naturally low-rank in real retrieval scenarios, making our approximation strategy both theoretically sound and empirically successful.

Takeaway. This work transforms bilinear retrieval from a theoretically appealing but computationally expensive method into a practical, scalable solution for real-world deployment. Our comprehensive analysis–covering 60,000 pointwise measurements–demonstrates that bilinear similarity models, when approximated via low-rank representations, maintain their expressive power while achieving dramatic efficiency gains. The singular value spectrum of the learned interaction matrix emerges as a powerful tool: It not only governs the trade-off between accuracy and compression, but also enables precise, theory-backed rank selection without the need for costly hyperparameter tuning.

Practically, this yields three critical benefits. First, **predictable performance**: our error bounds allow engineers to forecast retrieval accuracy before deployment. Second, **flexible deployment options**: we identify efficiency tiers–24x (ultra-efficient), 12x (balanced), and 8x (high-quality)–that meet varying application needs. Third, **theoretical validation**: the strong alignment between singular value decay and performance confirms that real retrieval systems benefit directly from our approximation framework.

6 Conclusion

This paper establishes bilinear similarities as a principled advancement for dense retrieval, providing the first rigorous theoretical foundation for understanding when and why sophisticated similarity functions outperform dot-product approaches. We prove that bilinear models are strictly more expressive under fixed embeddings, demonstrate concrete tasks where they achieve perfect performance while weighted dot-product models fail universally, and derive tight approximation bounds that enable efficient deployment. These theoretical advantages translate into substantial practical gains: up to 4x performance improvements on MS MARCO with efficient low-rank approximations achieving dramatic parameter compression (8x to 24x) while maintaining most benefits.

Our work transforms similarity function design from trial-and-error to principled engineering, challenging the dot-product orthodoxy that has dominated dense retrieval. The singular value spectrum emerges as a powerful tool for rank selection, enabling theory-guided deployment decisions. Most fundamentally, we demonstrate that sophisticated similarity modeling and computational efficiency are not mutually exclusive: Bilinear similarities offer a natural evolution of dual encoder architectures that captures essential feature interactions while remaining practically deployable. This reframes the traditional performance-efficiency trade-off and opens new possibilities for expressive yet scalable retrieval systems in an era of increasingly complex information needs.

References

1. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: Ms marco: A human generated machine reading comprehension dataset (2018), https://arxiv.org/abs/1611.09268
2. Bhojanapalli, S., Yun, C., Rawat, A.S., Reddi, S., Kumar, S.: Low-rank bottleneck in multi-head attention models. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
4. Dietz, L., Verma, M., Radlinski, F., Craswell, N.: Trec complex answer retrieval overview. In: TREC (2017)
5. Ding, B., Zhai, J.: Retrieval with learned similarities. In: Proceedings of the ACM on Web Conference 2025. p. 1626–1637. WWW '25, ACM (Apr 2025). https://doi.org/10.1145/3696410.3714822, http://dx.doi.org/10.1145/3696410.3714822
6. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. p. 55–64. CIKM'16, ACM (Oct 2016). https://doi.org/10.1145/2983323.2983769, http://dx.doi.org/10.1145/2983323.2983769
7. Hofstätter, S., Lin, S.C., Yang, J.H., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 113–122. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462891, https://doi.org/10.1145/3404835.3462891
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations (ICLR) (2022)
9. Hui, K., Yates, A., Berberich, K., de Melo, G.: PACRR: A position-aware neural IR model for relevance matching. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1049–1058. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1110, https://aclanthology.org/D17-1110/
10. Humeau, S., Shuster, K., Lachaux, M.A., Weston, J.: Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In: Proceedings of the 8th International Conference on Learning Representations (ICLR) (2020)
11. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised dense information retrieval with contrastive learning (2022), https://arxiv.org/abs/2112.09118
12. Kang, C., Liao, S., He, Y., Wang, J., Niu, W., Xiang, S., Pan, C.: Cross-modal similarity learning: A low rank bilinear formulation. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. p. 1251–1260. CIKM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806469, https://doi.org/10.1145/2806416.2806469

13. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550/
14. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 39–48. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401075, https://doi.org/10.1145/3397271.3401075
15. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017)
16. Liu, Z., Xiao, S., Shao, Y., Cao, Z.: RetroMAE-2: Duplex masked autoencoder for pre-training retrieval-oriented language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2635–2648. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.148, https://aclanthology.org/2023.acl-long.148/
17. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: Cedr: Contextualized embeddings for document ranking. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19, ACM (Jul 2019). https://doi.org/10.1145/3331184.3331317, http://dx.doi.org/10.1145/3331184.3331317
18. Mackenzie, J., MacAvaney, S., Goharian, N., Frieder, O.: Reneuir at sigir 2024: The third workshop on reaching efficiency in neural information retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3051–3054. ACM (2024)
19. Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.W., Yang, Y.: Large dual encoders are generalizable retrievers. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9844–9855. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.669, https://aclanthology.org/2022.emnlp-main.669/
20. Nogueira, R., Cho, K.: Passage re-ranking with bert (2020), https://arxiv.org/abs/1901.04085
21. Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., Cheng, X.: Text matching as image recognition (2016), https://arxiv.org/abs/1602.06359
22. Qian, Q., Baytas, I.M., Jin, R., Jain, A., Zhu, S.: Similarity learning via adaptive regression and its application to image retrieval (2015), https://arxiv.org/abs/1512.01728
23. Rendle, S.: Factorization machines. In: 2010 IEEE International Conference on Data Mining. pp. 995–1000 (2010). https://doi.org/10.1109/ICDM.2010.127
24. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In: Carpuat, M.,

de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3715–3734. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.272, https://aclanthology.org/2022.naacl-main.272/

25. Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.t., Smith, N.A., Zettlemoyer, L., Yu, T.: One embedder, any task: Instruction-finetuned text embeddings. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 1102–1121. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.71, https://aclanthology.org/2023.findings-acl.71/
26. Vasileiou, A., Eberle, O.: Explaining text similarity in transformer models (2024), https://arxiv.org/abs/2405.06604
27. Voorhees, E.M.: The trec robust retrieval track. In: ACM SIGIR Forum. vol. 39, pp. 11–20. ACM New York, NY, USA (2005)
28. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training (2024), https://arxiv.org/abs/2212.03533
29. Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: Proceedings of the 9th International Conference on Learning Representations (ICLR) (2021)
30. Zhai, J., Gong, Z., Wang, Y., Sun, X., Yan, Z., Li, F., Liu, X.: Revisiting neural retrieval on accelerators. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 5520–5531. KDD '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3580305.3599897, https://doi.org/10.1145/3580305.3599897